A novel two-tier virtualized PON transport to achieve ultra-low application-level latency in virtualized MESH-PON enabled MEC based Cloud-RAN

Sandip Das(1), Author 2(2), Author 3(2), Author 4(2), Author 5(1)

(1) CONNECT Centre, Trinity College Dublin, dassa@tcd.ie, marco.ruffini@scss.tcd.ie
(2) Intel Ireland, Ireland {emailID1, emailID2}@example.com

Abstract *(45 words max) word word word word word word word word word word word word word word word word word word word word word word word word word word word word word word word word word word word word word word word word word word word word.*

Introduction

Ultra-low application level end-to-end latency (between 1-10ms) is one of the key requirements for 5G and beyond in order to support latency-critical 5G applications such as Augmented Reality (AR), Intelligent Transport Systems (ITS), industry 4.0 etc[1]. Cloud Radio Access Networks (C-RAN), and Multi Access Edge Computing (MEC) are considered as the most promising technologies to support these requirements. However, three major bottlenecks in achieving such application-level low latency are the RAN access latency, fronthaul latency and transport to application latency (typically backhaul/internet).

In a traditional RAN access scheme, the RAN access latency is the amount of time the User Equipment (UE) application traffic needs to wait for allocation of uplink resource before transmission which is generally conveyed to UE via a set of downlink DCI messages in 5G NR. The latency between buffer status report by UE and the corresponding uplink resource grant allocation via DCI messages is 4 slot time (for e.g., 2 ms for 0.5 ms slot duration[2]). This is the largest contributing factor in RAN access latency. In order to address this bottleneck, there is an emerging concept of Coordinated Grant Scheduling (CGS) (for uplink) and Semi-Persistent Scheduling (SPS) (in downlink) which semi-statically pre-allocates uplink resources (typically a group of Physical Resource Blocks (PRBs)) to UEs in order to send their uplink traffic without requesting and therefore waiting for the uplink resources.

Another important source of delay occurs in the fronthaul when the RAN cell works on a eCPRI split (i.e., 7.2 is typical) that operates over a Passive Optical Network (PON). Here, if the PON and RAN schedules are not coordinated, data from the UE will also need to queue at the ONU side waiting for the PON grant to be provided by the OLT. This coordination issues was recently solved with the development of a new Coordinated Transport Interface (CTI)[3], which requires the OLT to fetch prior UE uplink scheduling information from DU/CU and use it to calculate the RU-ONU arrival packet size and start time to determine the uplink grant and convey it to RU-ONUs beforehand in order to remove the queuing of packets at the ONUs. However, as the allocation of CGS is typically done through RRC semi-statically, and the traffic on the corresponding CGS resources can not be reported in advance, the OLT has no-way of absolutely determining the CGS resources and the amount of traffic carried over in those resources. This would incur additional queuing latency with CO-DBA if the allocated CGS resource and the traffic over it is under-estimated. Therefore, the CTI interface and the corresponding DBA needs to be updated to process this information properly to achieve a ultra-low RU-DU fronthaul transport latency also with RAN access latency significantly reduced via CGS.

The third source of delay occurs when the Distributed Unit (DU) processed data is sent to the application processing. Typically, this is sent over the midhual network to either at the Central Office (CO) or at another MEC depending upon application-latency requirements. This midhaul network usually consists of active optical network elements, thus can have significant latency depending on the route and the traffic load at the midhaul network. Therefore, In order to address these shortcomings, we propose the following contributions in this this article:

1. We propose a two-tier joint vPON transport method to achieve ultra-low application-level latency in MESH-PON enabled MEC based Cloud-RAN.
2. In the first tier, we propose an enhanced coordinated DBA with CGS to jointly achieve ultra-low the fronthaul latency and RAN access latency.
3. The proposed second-tier vPON transports the DU processed traffic directly to the application (hosted at a different MEC) without the need transporting the traffic via OLT backplane and midhaul network. Therefore achieving an overall ultra-low end-to-end latency at the application level.

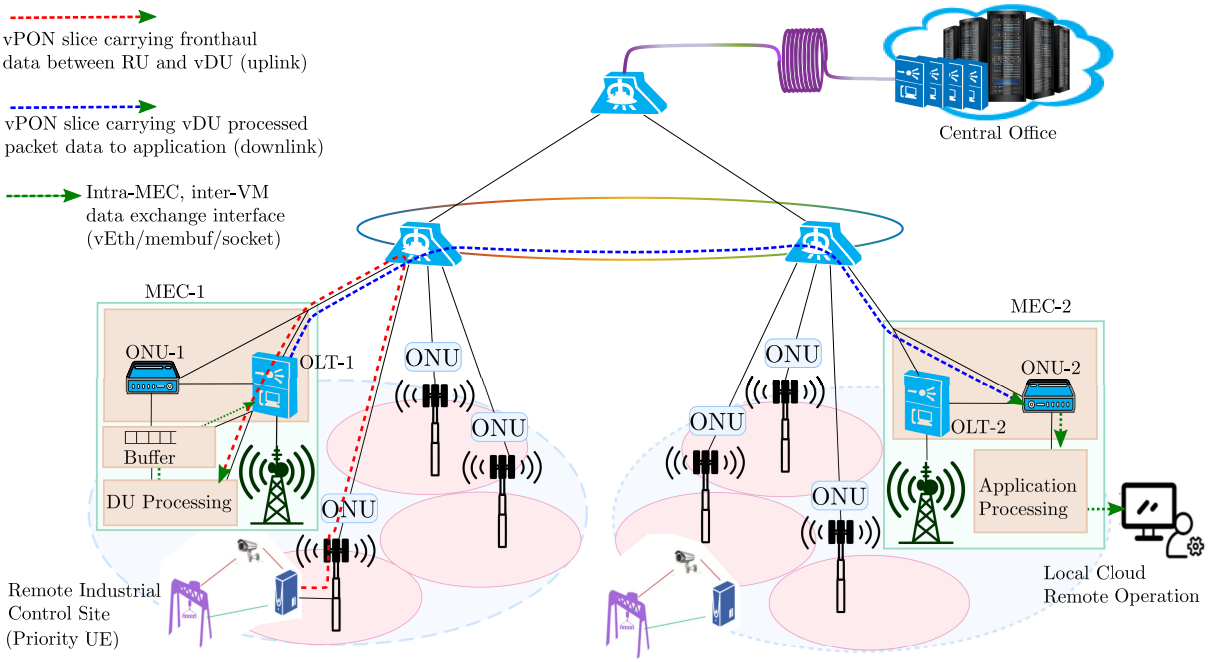

Fig. 1: System architecture for our Variable Bandwidth Fronthaul approach to cloud-RAN

System Architecture and the Proposed two-tier joint vPON scheduling method

Fig. 1 presents the system architecture and use case. We consider a fixed-wireless converged architecture, where a TWDM-PON is used for sharing C-RAN fronthaul with residential broadband. At the application level, we consider the use case of remote industrial control for which the application level ultra-low latency ($<$1ms or 1-5ms depending upon the application) is one of the most crucial requirement[?]. In order to achieve this tight application-level low latency, we propose the following tow-tier joint vPON scheduling method that targets the above mentioned three latency components. In the following, we describe the details of our architecture:

In order to achieve the ultra-low latency in the fronthaul part of the system, the TWDM-PON architecture can be enhanced with our previously proposed virtualized MESH-PON architecture[4]. Further, MEC nodes with limited processing capacity can be deployed in the macrocell site, where the CU/DU processing can be virtualized (vCU/vDU). RU-ONUs servicing the traffic of remote industrial control (we refer it priority-UE traffic hereinafter) can create a virtual PON slice with the OLT at the nearby MEC nodes (MEC-1 in this case) to transport the fronthaul data at an ultra-low latency. We refer this vPON slice as the 1^{st}-tier vPON slice and it's path is illustrated with red-colored dotted line in the Fig. 1.

The first-tier vPON slice in our proposed architecture achieves the ultra-low latency in the fronthaul and the RAN access part of the system. We propose the following enhancement of CO-DBA in the first-tier vPON slice to efficiently incorporate the CGS resources to jointly reduce RAN access and fronthaul latency. As priority-UE's application has SLA for low-latency, the RRC module of 5G-NR semi-statically allocates a set of CGS resources which UE can acquire to send uplink traffic whenever it has application level packet to be transported. On the other hand, RRC also passes the information of the allocated CGS resources to the OLT to incorporate this CGS information while calculating uplink grants for CO-DBA. There are several ways to achieve this, one relatively simple and conservative approach is to include the CGS resources into account while calculating the uplink grants regardless of how many PRBs of the allocated CGS resources the UE is actually occupied depending on the current traffic. Another way is to estimate the the actual UE application and consequently estimate the percentage of the CGS resources that is occupied in the uplink and use it along with the typical mac scheduling information of calculating the grants. Use of CGS resource for the uplink transmission achieves ultra-low RAN access latency while the enhanced CO-DBA ensures the corresponding uplink grants to be properly allocated by incorporating the CGS resources in order to avoid queuing of fronthaul packets due to variation of the fronthaul rates due to the UE's traffic carried in the CGS resources.

The second-tier vPON slice in our proposed architecture targets to achieve ultra-low latency in the midhaul where the CU/DU processed data packets are transported to application residing at a cloud processing unit. In order to achieve ultra-low latency, it is fair to assume that the application also residing at a nearby MEC location (shown as MEC-2 in Fig. 1). Ideally in a typical PON

based fronthaul/midhaul/backhaul deployment, in order to transport the CU/DU processed data from MEC-1 to application at MEC-2, the data has to travel all the way to CO, OEO conversion and back to MEC-2 via OLT-2. Instead, we propose to transport this CU/DU processed data to application at MEC-2 directly over the EAST-WEST link[1] using our proposed second-tier vPON transport. This can be done in two alternative ways. the first possible way is to configure the vPON slice of OLT-1 (at MEC-1) to include the ONU at the MEC-2 and send the traffic to MEC-2 at the application over the next downlink period of the same vPON slice. We refer this vPON slice as the 2^{nd}-tier vPON slice and it's path is illustrated with blue-colored dotted line in the Fig. 1. This when coordinated with the first vPON slice, the combined method is hereinafter referred as UL-DL (Uplink-Downlink) method.

Another alternative is by configuring the vPON slice of OLT-2 (at MEC-2) to temporarily include the ONU at MEC-1 and send the traffic to MEC-2 using the uplink of the vPON slice of OLT-2. This when coordinated with the first vPON slice, the combined method is hereinafter referred as UL-DL (Uplink-Downlink) method. In this limited scope of this manuscript, we consider only the first alternative option of the above two (i.e., downlink on the second tier vPON transport). However, it is also worth exploring the other alternative (i.e, uplink over the second vPON slice) as there can be possible scenarios when downlink transport in second-tier is not feasible due to lack of downlink bandwidth resources for the OLT at MEC-1. This would then require the use of low-latency uplink transport over vPON slice-2 to send CU/DU data to the application.

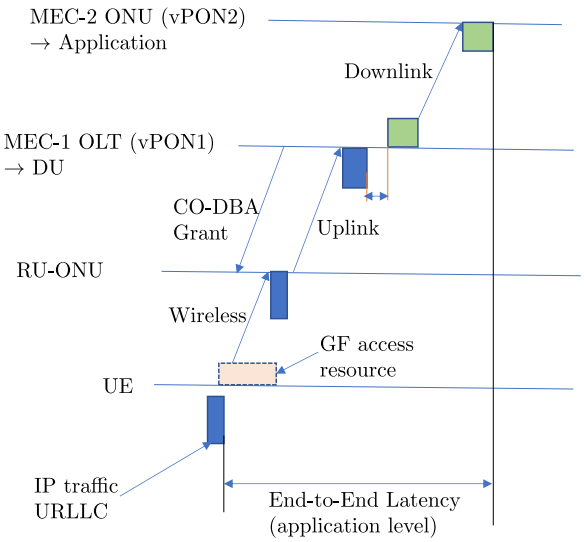

Fig. 2: The two-tier DBA protocol for facilitating the application-level low-latency in MESH-PON enabled MEC based Cloud-RAN

The two-tier DBA protocol for facilitating the application-level end-to-end low-latency is demonstrated in Fig. 2. We consider the application level IP traffic from the UE which requires ultra-low end-to-end latency at the application level. Considering the UL-DL method of two-tier joint vPON transport scheme, the application traffic from the UE (remote industrial site) uses Grant Free (GF) resource (allocated CGS resource) to transmit their uplink data over RAN. This reduces the wireless RAN access delay significantly. This is then transported over fronthaul to MEC-1 for CU/DU processing using the first-tier vPON transport in uplink (shown as red-dotted path). Here, our proposed enhanced CO-DBA utilizes the allocated CGS resource information to update the CO-DBA grants in order to avoid the queuing of fronthaul packets due to the traffic carried over the CGS resources. After the CU/DU processing, the packets are handed over in the downlink OLT queue via internal shared memory interface or vEth interface. The OLT at MEC-1 then transports those packets to ONU at MEC-2 as a second-tier vPON transport at the immediate-next downlink interval. The downlink received packets are finally sent to the application using common-buffer interface.

Simulation Model

Results

Conclusion

Acknowledgment

Financial support from SFI 15/US-C2C/I3132, 14/IA/2527 and 13/RC/2077 & NSF is gratefully acknowledged.

References

[1] "Verticals URLLC Use Cases and Requirements". NGMN Allaince, Feb 2020.

[2] E. Dahlman, S. Parkvall, and J. Sköld, "5G NR: the next generation wireless access technology, Chapter 14: Scheduling". Elsevier, Academic Press, 2021.

[3] O-RAN Fronthaul Working Group 4, "Cooperative Transport Interface Transport Control Plane Specification," O-RAN alliance, Mar. 2021.

[4] S. Das, F. Slyne, and M. Ruffini. "Optimal Slicing of Virtualised Passive Optical Networks to Support Dense Deployment of Cloud-RAN and Multi-Access Edge Computing., IEEE Network (to appear), Mar 2022, arXiv preprint arXiv:2203.11857.

[5] Common Public Radio Interface (CPRI); Interface Specification. in Specification CPRI, July 2014.