Dynamic RB scheduling for different slices of eMBB and URLLC in the O-RAN system

***Abstract*—**

I. Introduction

II. System Model and Problem Formulation

A. System Model

Assume we have two different services, namely, eMBB and URLLC; The system is serving a set of $\mathcal{U}_1$ eMBB single-antenna user equipments (UEs) and a set of $\mathcal{U}_2$ URLLC UEs. Assume our system consists of K, preallocated physical resource blocks (PRBs). Moreover, the system considers to have M_s^d VNFs for the processing of O-DU, M_s^c VNFs for the processing of O-CU-UP of eMBB and URLLC ($s \in \{1, 2\}$). Virtual network functions (VNFs) are functional blocks of the system. Each VNF instance runs on a virtual machine (VM) using resources from the data centers. Moreover, we assume there is a cell with one multi-antenna O-RU that serves UEs. Furthermore, we suppose the channel distribution is known.

B. The Achievable Rate

The eMBB services typically use more than a one-time slot. But URLLC services use part of a time slot (mini-slot) since it has short packet transmission that it is indicated in Fig 2. In addition, the URLLC must be punctured as soon as it has requested service as it requires very low latency. Our current work allocates RBs to eMBB users at the beginning of each time slot using the PF principle, a scheduling strategy that balances throughput with fairness [1].

The achievable data rate and the expectation of the achivable data rate for the i^{th} UE request eMBB slice can be written as $\mathcal{R}_i^e(t)$, and $\bar{\mathcal{R}}_i^e(t)$, respectively.

$$\begin{aligned}\mathcal{R}_i^e(t) &= \sum_{k=1}^{K} e_i^k(t) B(1 - n_i^k(t)) \log_2(1 + \frac{p_i^k(t) h_i^k(t)}{B \times N_0}),\\ \bar{\mathcal{R}}_i^e(t) &= \mathbb{E}_h[\mathcal{R}_i^e(t)],\end{aligned} \tag{1}$$

where B is the bandwidth of RBs. Also, $B \times N_0$ denotes the power of Gaussian additive noise. Moreover, $e_i^k(t) \in \{0, 1\}$ is a binary variable that illustrates whether PRB k is assigned to the i^{th} eMBB UE or not. $p_i^k(t)$ represents the transmission power allocated by O-RU to i^{th} UE of eMBB using PRB k. $h_i^k(t)$ is the channel gain of a wireless link from O-RU to the i^{th} eMBB UE using k^{th} PRB which is Rayleigh fading. Furthermore, $n_i^k(t)$ is the percentage of RB k using eMBB i that is punctuated by URLLC UEs is denoted as follows,

$$n_i^k(t) = \psi_i^k(t) \frac{\sum_{j \in \mathcal{U}_2} \zeta_j^u(t) \nu^u \tau}{S_{max}(t) \tau}, \tag{2}$$

Where, $\psi_i^k(t)$ is the probability of puncturing eMBB UE i using RB k. Furthermore, $\zeta^u(t)$ is the arrival rate of URLLC UEs (arrival/slot/user). Moreover, $\mathcal{U}_2$ is the set of URLLC UEs in the system. In addition, ν^u is the URLLC packet size, and τ is the number of slot per second (slot/sec). Assume the packet arrival of URLLC UEs follows a Poisson process with arrival rate $\lambda(t)$. Therefore, the arrival data rate of URLLC j is $\lambda_j(t) = \zeta_j^u(t) \nu^u \tau$. Moreover, $S_{max}(t)$ is the flow density refers to how much data the system can handle for all URLLC users at any given time (bits/slot). Moreover, we have $\sum_i e_i^k(t) \leq 1$ to guarantee that each RB is allocated to a maximum of one eMBB UE.

Since the blocklength in URLLC is finite, the achievable data rate for the j^{th} UE request in the URLLC service, is not achieved from Shannon Capacity formula. So, for the short packet transmission the achievable data rate and its expectation is written as follow, respectively,

$$\begin{aligned}\mathcal{R}_j^u(t) &= \sum_{i=1}^{U_1} \sum_{k=1}^{K} m_i^k(t) B \log_2(1 + \frac{p_j^k(t) h_j^k(t)}{B \times N_0}) - \zeta_j^k(t)),\\ \bar{\mathcal{R}}_j^u(t) &= \mathbb{E}_h[\mathcal{R}_j^u(t)],\end{aligned} \tag{3}$$

where $\zeta_j^k(t) = \log_2(e) Q^{-1}(\epsilon) \sqrt{\frac{C_j^k(t)}{N_j^k(t)}})$ where ϵ is the transmission error probability, Q^{-1} is the inverse of Q function (i.e., Gaussian), $C_j^k(t) = 1 - \frac{1}{(1+\rho_j^k(t))^2}$ depicts the channel dispersion of UE j of URLLC, puncturing mini-slots of PRB k and $N_j^k(t)$ represents the blocklength of it. Moreover, $\rho_j^k(t) = \frac{p_j^k(t) h_j^k(t)}{B \times N_0}$ is the SNR of UE j in URLLC service. Also, $m_i^k(t)$ is denoted as follow

$$m_i^k(t) = \frac{e_i^k(t) n_i^k(t)}{|\mathcal{U}_2|}, \tag{4}$$

where, $|\mathcal{U}_2|$ is the number of URLLC UEs in the system.

C. Mean Delay

In this part, the mean processing delay for each service is obtained. Suppose the mean total processing delay is depicted as T_{tot},

$$T^{\text{tot}} = T^{RU} + T^{\text{proc}}, \tag{5a}$$

$$T^{\text{proc}} = T^{DU} + T^{CU}. \tag{5b}$$

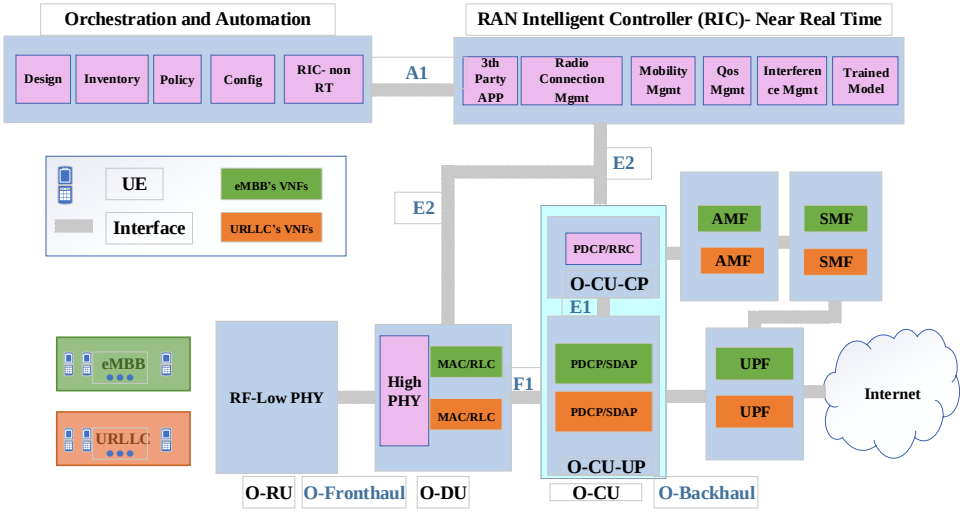

Fig. 1: Network sliced ORAN system

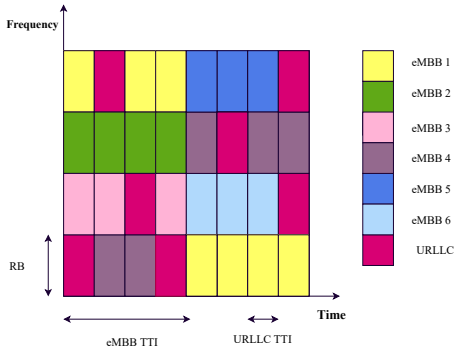

Fig. 2: RB scheduling

We assume that the packet arrival rate of URLLC UEs follows a Poisson process with arrival rate $\lambda_j(t)$ for the j^{th} UE. Thus, we have $\lambda_j(t) = \zeta_j^u(t)\nu^u\tau$. Therefore, the mean arrival data rate of the O-CU-UP layer is $\alpha^C(t) = \sum_{j\in\mathcal{U}_2}\lambda_j(t)$. Assume the mean arrival data rate for URLLC slice (α) is approximately equal to the mean arrival data rate of the the O-DU (α^D). so $\alpha(t) = \alpha^C(t) \approx \alpha^D(t)$, Because the amount of data traffic transferred along the route (regardless of frame changes) is constant. Since, by using Burke's theorem, the mean arrival data rate of the second layer, which are processed in the first layer, is still poisson with rate α. It is assumed that there are load balancers in each layer for each service to divide the incoming traffic to VNFs equally. Suppose the baseband processing of each VNF is depicted as M/M/1 processing queue. Each packet is processed by one of the VNFs of a slice. So, the mean delay for the URLLC slice in the O-DU,and the O-CU is modeled as M/M/1 queue, is formulated as follows, respectively [2]–[4],

$$\begin{aligned} T^{DU} &= \frac{1}{\mu^d - \alpha(t)/M^d(t)}, \\ T^{CU} &= \frac{1}{\mu^c - \alpha(t)/M^c(t)}, \end{aligned} \tag{6}$$

where $M^d(t)$,and $M^c(t)$ are the variables that depict the number of VNFs in O-DU,and O-CU-UP, respectively. Moreover, $1/\mu^d$, and $1/\mu^c$ are the mean service time of the O-DU, and O-CU layers, respectively. Besides, α is the arrival rate which is divided by load balancer before arriving to the VNFs. The arrival rate of each VNF in each layer for URLLC slice is α/M^l $l \in \{d, c\}$.

Suppose the mean transmission delay of the j^{th} UE of the URLLC service on the wireless link is denoted by $T_j^{RU}(t)$. Assume the arrival data rate of wireless link for each UE j of URLLC service is $\lambda_j(t)$ As a result, we have $\sum_{j\in\mathcal{U}_2}\lambda_j(t) = \alpha(t)$. Moreover, The service time of transmission queue for UE j requesting URLLC service has an exponential distribution with mean $1/R_j^u(t)$ and can be modeled as a M/M/1 queue [2]–[4].

Therefore, the mean delay of the transmission layer for UE j in URLLC slice is

$$T_j^{RU}(t) = \frac{1}{\mathcal{R}_j^u(t) - \lambda_j(t)}. \tag{7}$$

D. Reliability of URLLC

As we know, UEs request URLLC services, require services with low latency. For the M/M/1 system, the probability of the delay for URLLC service in the O-RU is as follow,

$$P_r\{T_j^{RU} \geq T_{RU}^{max}\} = e^{-(R_j^u - \alpha)T_{RU}^{max}} \tag{8}$$

Also, we do not consider the reliability for O-CU and O-DU.

E. Problem Statement

In this paper, we strive to maximize the sum rate of all eMBB UEs and minimize the delay of URLLC UEs while imposing constraints on their performance based on their QoS. The optimization problem is formulated as follow,

$$\max_{\boldsymbol{E},\boldsymbol{\Psi},\boldsymbol{M}} \sum_i \bar{\mathcal{R}}_i^e(t) - \eta \sum_j T^{\text{tot},u}(t) \tag{9a}$$

$$\text{subject to} \quad \bar{\mathcal{R}}_i^e(t) \geq \mathcal{R}_{min}^e \quad \forall i \in \mathcal{U}_1, \tag{9b}$$

$$Pr\{\bar{\mathcal{R}}_i^e(t) \leq R_{min}^e\} \leq \epsilon, \quad \forall i \in \mathcal{U}_1, \tag{9c}$$

$$T_j^{\text{tot},u}(t) \leq T_{min}^u \quad \forall j \in \mathcal{U}_2, \tag{9d}$$

$$Pr\{T_j^{\text{RU},u}(t) \geq T_{min}^u\} \leq \epsilon \quad \forall j \in \mathcal{U}_2, \tag{9e}$$

$$\sum_{i=1}^{U_1} e_i^k(t) \leq 1 \quad \forall k \in \{1, ..., K\}, \tag{9f}$$

$$e_i^k(t) \in \{0, 1\} \quad \forall i, k, \tag{9g}$$

$$0 \leq n_i^k(t) \leq 1 \quad \forall i, k, \tag{9h}$$

$$\mu^l \geq \alpha / M^l \quad l \in \{c, d\}, \tag{9i}$$

$$\bar{\mathcal{R}}^u(t) \geq \lambda_j^u(t) \quad \forall j \in \mathcal{U}_2, \tag{9j}$$

$$0 \leq M^l \leq M_{max}^l \quad l \in \{c, d\}, \tag{9k}$$

Where, (10b), guarantees the minimum data rate of eMBB UEs, also (9c), supports the reliability of eMBB while puncturing the URLLC. In addition, (9d), and (9e) guarantee the latency and reliability of URLLC UEs, respectively. Furthermore, eMBB RB allocation constraint is indicated by (9f), and (9g). (9h), indicate that the punctured mini-time slot are fewer than the total number of mini-slots in the RB. (9i) and (9j) denotes the stability of the M/M/1 queue model. (9k) restricts the number of VNF in each slice due to the limited resources.

III. Proposed Algorithm

In this section, we talk about our proposed algorithm to solve the problem (9). Since this problem is mixed-integer nonlinear programming (MINLP) with binary and integer variables, it is complicated to solve.

We can solve this problem on a two-time scale. On a large time scale, the problem of assigning RB to eMBB UEs is solved, Also, in this time scale, we can estimate the optimal number of VNFs, and in the small time scale, the problem of URLLC puncturing is solved.

A. Large time scale

In this time scale, we want to solve the problem of eMBB scheduling and finding the optimal number of VNF for URLLC UEs. Here, we suppose the puncturing of URLLC is fixed, hence, we want to solve the problem (9). The problem (9), is altered to the following problem

$$\max_{\boldsymbol{E},\boldsymbol{M}} \sum_i \bar{\mathcal{R}}_i^e(t) - \eta T^{\text{proc},u}(t) \tag{10a}$$

$$\text{subject to} \quad \bar{\mathcal{R}}_i^e(t) \geq \mathcal{R}_{min}^e \quad \forall i \in \mathcal{U}_1, \tag{10b}$$

$$Pr\{\bar{\mathcal{R}}_i^e(t) \leq R_{min}^e\} \leq \epsilon, \quad \forall i \in \mathcal{U}_1, \tag{10c}$$

$$\sum_{i=1}^{U_1} e_i^k(t) \leq 1 \quad \forall k \in \{1, ..., K\}, \tag{10d}$$

$$e_i^k(t) \in \{0, 1\} \quad \forall i, k, \tag{10e}$$

$$0 \leq n_i^k(t) \leq 1 \quad \forall i, k, \tag{10f}$$

$$\mu^l \geq \alpha / M^l \quad l \in \{c, d\}, \tag{10g}$$

$$0 \leq M^l \leq M_{max}^l \quad l \in \{c, d\}, \tag{10h}$$

Generally, let's assume O-CU and O-DU use the same processor. Consequently, the formulation becomes simpler. Despite this, the formulation remains the same, and the problem can still be solved similarly. Consequently, we have $\mu = \mu^c \approx \mu^d$. Additionally, the mean arrival data rate for the O-DU layer (α^C) is the same as the O-CU-UP layer (α^C). So $\alpha = \alpha^C \approx \alpha^D$. Accordingly, we can have $M^d(t) = M^c(t)$. Therefore, $T^{DU} = T^{CU}$, and $T^{proc} = 2 \times T^{DU}$.

The problem (10), is still mixed integer non-linear programming (MINP). Here, the problem (10) can also be decomposed into two sub-problems which is depicted as follow.

1) sub-problem 1: In this section, we want to solve the sub-problem of RB allocation of eMBB UEs. The problem can be written as follow.

$$\max_{\boldsymbol{E}} \sum_i \bar{\mathcal{R}}_i^e(t) \tag{11a}$$

$$\text{subject to} \quad \bar{\mathcal{R}}_i^e(t) \geq \mathcal{R}_{min}^e \quad \forall i \in \mathcal{U}_1, \tag{11b}$$

$$\sum_{i=1}^{U_1} e_i^k(t) \leq 1 \quad \forall k \in \{1, ..., K\}, \tag{11c}$$

$$e_i^k(t) \in \{0, 1\} \quad \forall i, k, \tag{11d}$$

2) sub-problem 2: In this section, we want to solve the sub-problem of finding the optimal number of VNF for URLLC services. The problem is formulated as follow.

$$\min_{\boldsymbol{M}} T^{\text{proc},u}(t) \tag{12a}$$

$$\text{subject to} \quad \mu^l \geq \alpha / M^l \quad l \in \{c, d\}, \tag{12b}$$

$$0 \leq M^l \leq M_{max}^l \quad l \in \{c, d\}, \tag{12c}$$

B. Small time scale

In the small time scale, we assume that the eMBB RB allocation is performed and the optimal number of VNF for URLLC is obtained. Therefore, the problem of puncturing URLLC UEs is performed.

$$\min_{\boldsymbol{\Psi}} \sum_j T^{RU,u}(t) \tag{13a}$$

$$\text{subject to} \quad T_j^{\text{tot},u}(t) \leq T_{min}^u \quad \forall j \in \mathcal{U}_2, \tag{13b}$$

$$Pr\{\bar{\mathcal{R}}_i^e(t) \leq R_{min}^e\} \leq \epsilon, \quad \forall i \in \mathcal{U}_1, \tag{13c}$$

$$Pr\{T_j^{\text{RU},u}(t) \geq T_{min}^u\} \leq \epsilon \quad \forall j \in \mathcal{U}_2, \tag{13d}$$

$$0 \leq n_i^k(t) \leq 1 \quad \forall i, k, \tag{13e}$$

$$\bar{\mathcal{R}}^u(t) \geq \lambda_j^u(t) \quad \forall j \in \mathcal{U}_2, \tag{13f}$$

REFERENCES

[1] B. Shi, F. Zheng, C. She, J. Luo, and A. G. Burr, "Risk-resistant resource allocation for embb and urllc coexistence under m/g/1 queueing model," *IEEE Transactions on Vehicular Technology*, 2022.

[2] J. Tang, W. P. Tay, T. Q. Quek, and B. Liang, "System cost minimization in cloud RAN with limited fronthaul capacity," *IEEE Transactions on Wireless Communications*, vol. 16, no. 5, pp. 3371–3384, 2017.

[3] P. Luong, F. Gagnon, C. Despins, and L.-N. Tran, "Joint virtual computing and radio resource allocation in limited fronthaul green C-RANs," *IEEE Transactions on Wireless Communications*, vol. 17, no. 4, pp. 2602–2617, 2018.

[4] P. Luong, C. Despins, F. Gagnon, and L.-N. Tran, "A novel energy-efficient resource allocation approach in limited fronthaul virtualized C-RANs," in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*. IEEE, 2018, pp. 1–6.