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

Consider two different services of eMBB and URLLC in O-RAN architecture; each offers various applications. Therefore, we assume two different slices for eMBB and URLLC services. The system is serving a set of $\mathcal{U}_e$ eMBB single-antenna user equipment(UE) and a set of $\mathcal{U}_u$ URLLC UE. Suppose we have K preallocated resource blocks (RBs) for serving these two services. Moreover, the system considers having virtual network function (VNFs) in O-DU and O-CU to process packets. Each service has its VNFs for processing due to the isolation of slices. Let the system 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 services ($s \in \{e, u\}$ for eMBB and URLLC respectively). A VNF is a system's core component. Every VNF instance is hosted on a virtual machine (VM) using resources from the data centers. Moreover, we assume there is a cell with one O-RU that serves UEs. Furthermore, we suppose the channel distribution is known. The eMBB services generally use more than a one-time slot. But URLLC services use part of a time slot (mini-slot) since it has a short packet transmission. By using the PF principle, we currently allocate RBs at the beginning of each eMBB time slot, a scheduling strategy that balances throughput and fairness [1]. However, puncturing the URLLC once requested service is necessary as it requires very low latency.

B. The Achievable Rate

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

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

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 RB k is assigned to the ith 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 ith eMBB UE using k^{th} PRB which is Rayleigh fading. Furthermore, $n_i^k(t)$ denotes the percentage

of RB k using eMBB i that is punctured by the set of URLLC service is indicates as follows,

$$n_i^k(t) = \frac{x_i^k(t)}{L} \tag{2}$$

where, $x_i^k(t)$ is the number of punctured mini-slot of RB k that is used by eMBB UE i. Also, L is the number of mini-slots in each RB [2].

Moreover, we have $\sum_i e_i^k(t) \le 1$ 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 jth 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,

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

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 jth UE of URLLC service, puncturing mini-slots of RB 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

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

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

C. Mean Delay

In this section, the mean processing delay for each service is determined. Let us suppose the mean total processing time is

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

We assume that the packet arrival rate of each UE i in each service s follows a Poisson process with arrival rate $\lambda_{i,s}(t)$. Therefore, the mean arrival data rate of the O-CU-UP layer is $\alpha_s^C(t) = \sum_{i \in \mathcal{U}_s} \lambda_{i,s}(t)$, $s \in \{e, u\}$ which is for eMBB and URLLC service, respectively. Assume the mean arrival data rate for URLLC slice (α) is approximately equal to the mean arrival data rate of the O-DU (α^D). so

$\alpha_s(t) = \alpha_s^C(t) \approx \alpha_s^D(t), s \in \{e, u\}$. 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 each slice ($s \in \{e, u\}$ which is for eMBB and URLLC service, respectively) in the O-DU,and the O-CU is modeled as M/M/1 queue, is formulated as follows, respectively [3]–[5],

$$T_s^{DU} = \frac{1}{\mu_s^d - \alpha_s(t)/M_s^d(t)},$$
$$T_s^{CU} = \frac{1}{\mu_s^c - \alpha_s(t)/M_s^c(t)}, \tag{6}$$

where $M_s^d(t)$,and $M_s^c(t)$ are the variables that depict the number of VNFs in O-DU,and O-CU-UP, respectively. Moreover, $1/\mu_s^d$, and $1/\mu_s^c$ are the mean service time of the O-DU, and O-CU layers, respectively. Besides, α_s 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 each slice is α_s/M_s^l $l \in \{d, c\}$.

Suppose the mean transmission delay of the ith UE of the service s ($s \in \{e, u\}$ which is for eMBB and URLLC service, respectively) on the wireless link is denoted by $T_{i,s}^{RU}(t)$. Assume the arrival data rate of wireless link for each UE i of service s is $\lambda_{i,s}(t)$ As a result, we have $\sum_{i \in \mathcal{U}_s} i, s(t) = \alpha_s(t), s \in \{e, u\}$. Moreover, The service time of transmission queue for UE i requesting service s has an exponential distribution with mean $1/R_i^s(t)$ and can be modeled as a M/M/1 queue [3]–[5].

Therefore, the mean delay of the transmission layer for UE i in slice s is

$$T_{i,s}^{RU}(t) = \frac{1}{\mathcal{R}_i^s(t) - \lambda_{i,s}(t)}. \tag{7}$$

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,

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

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,

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

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

$$\sum_k^K \sum_i^{U_e} p_i^k(t) \leq P_{\max}, \tag{9c}$$

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

$$T_i^{\text{tot},e}(t) \leq T_{min}^e \quad \forall i \in \mathcal{U}_e, \tag{9e}$$

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

$$\mathcal{R}_j^u(t) \geq \mathcal{R}_{min}^u \quad \forall j \in \mathcal{U}_u, \tag{9g}$$

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

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

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

$$x_i^k(t) \in \{0, 1, ..., L\} \quad \forall i, k, \tag{9k}$$

$$\mu_s^l \geq \alpha_s/M_s^l \quad l \in \{c, d\}, s \in \{e, u\}, \tag{9l}$$

$$\mathcal{R}_i^s(t) \geq \lambda_{i,s}(t) \quad \forall i \in \mathcal{U}_s, s \in \{e, u\} \tag{9m}$$

$$0 \leq M_e^l + M_u^l \leq M_{max}^l \quad l \in \{c, d\}, \tag{9n}$$

$$0 \leq M_e^l, M_u^l \quad l \in \{c, d\}, \tag{9o}$$

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

III. PROPOSED ALGORITHM

In this section, we are talking 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 NP-hard and complicated to solve. Also, this is a two-time scale problem. Therefore, We need to solve this problem on a two-time scale. On a large time scale, the problem of assigning RB to eMBB service is obtained and then the power allocation is performed. Moreover, in this time scale, we can estimate the optimal number of VNFs for each service, 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, power allocation, and finding the optimal number of VNF for each service. 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

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

$$\text{subject to} \quad (9b), (9c), (9d), (9e), (9i), (9j), (9l), \quad (10b)$$

$$\mathcal{R}^e(t) \geq \lambda_{i,e}(t) \quad \forall i \in \mathcal{U}_e, \quad (10c)$$

$$(9n), (9o), \quad (10d)$$

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 be decomposed into two sub-problems which is depicted as follow. In this first sub-problem, we assume that the channel gain is not known, and we aim to obtain the optimal number of VNFs and RB allocation to the eMBB service. In the second sub-problem, we know the channel gain with error ($h = h + \Delta h$), and we want to obtain the power allocation.

1) eMBB RB Allocation and VNF Assignment: In this section, we want to solve the sub-problem of RB allocation of eMBB service and find the optimal number of VNFs for each service. The problem can be written as follow.

$$\max_{\boldsymbol{E}, \boldsymbol{M}} \sum_i \bar{\mathcal{R}}_i^e(t) - T^{\text{proc},u}(t) \quad (11a)$$

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

$$\bar{T}_i^{\text{tot},e}(t) \leq T_{min}^e \quad \forall i \in \mathcal{U}_e, \quad (11c)$$

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

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

$$\mu_s^l \geq \alpha_s / M_s^l \quad l \in \{c, d\}, s \in \{e, u\}, \quad (11f)$$

$$0 \leq M_e^l + M_u^l \leq M_{max}^l \quad l \in \{c, d\}, \quad (11g)$$

$$0 \leq M_e^l, M_u^l \quad l \in \{c, d\}, \quad (11h)$$

The problem (11), can be decomposed into two subproblems using primal decomposition. $\boldsymbol{E}$ is the private variable and $\boldsymbol{M}$ is the decoupling variable. Therefore, the first subproblem is

$$\max_{\boldsymbol{E}, M_e} \sum_i \bar{\mathcal{R}}_i^e(t) \quad (12a)$$

$$\text{subject to} \quad (11b), (11c), (11d), (11e), \quad (12b)$$

$$\mu_e^l \geq \alpha_e / M_e^l \quad l \in \{c, d\}, \quad (12c)$$

$$0 \leq M_e^l \leq \tau \quad (12d)$$

Moreover, the second subproblem is

$$\min_{M_u} T^{\text{proc},u}(t) \quad (13a)$$

$$\text{subject to} \quad \mu_u^l \geq \alpha_s / M_u^l \quad l \in \{c, d\}, \quad (13b)$$

$$0 \leq M_u^l \leq M_{max}^l - \tau \quad (13c)$$

2) eMBB Power Allocation: In this section, we want to solve the sub-problem of power allocation for eMBB UEs. The problem is formulated as follow.

$$\max_{\boldsymbol{P}} \sum_i \mathcal{R}_i^e(t) \quad (14a)$$

$$\text{subject to} \quad (9c), (9d), (9e), \quad (14b)$$

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 considered.

$$\min_{\boldsymbol{X}} \sum_j T^{RU,u}(t) \quad (15a)$$

$$\text{subject to} \quad Pr\{\mathcal{R}_i^e(t) \leq R_{min}^e\} \leq \epsilon_2, \quad \forall i \in \mathcal{U}_e, \quad (15b)$$

$$Pr\{T_j^{\text{tot},u}(t) \geq T_{min}^u - T^{\text{proc}}\} \leq \epsilon_3 \quad \forall j \in \mathcal{U}_u, \quad (15c)$$

$$x_i^k(t) \in \{0, 1, ..., L\} \quad (15d)$$

$$\mathcal{R}_j^u(t) \geq \lambda_j^u(t) \quad \forall j \in \mathcal{U}_u, \quad (15e)$$

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] M. Alsenwi, N. H. Tran, M. Bennis, S. R. Pandey, A. K. Bairagi, and C. S. Hong, "Intelligent resource slicing for eMBB and URLLC coexistence in 5G and beyond: A deep reinforcement learning based approach," *IEEE Transactions on Wireless Communications*, 2021.

[3] 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.

[4] 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.

[5] 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.