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 preallocated slices serving two services, eMBB, and URLLC services; eMBB Service consists of U_1 single-antenna user equipments (UEs) and URLLC service consists of U_2 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.

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. 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 for the i^{th} UE request eMBB slice can be written as $\mathcal{R}_i^e(t)$.

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

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. In addition, S is the total number of mini-slots of a PRB. Furthermore, $n_i^k(t)$ is the number of URLLC punctured slots that is denoted as follow,

$$n_i^k(t) = \sum_{j=1}^{U_2} \zeta_j(t) \tau \mathfrak{N} \tag{2}$$

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 is approximated as follow

$$\mathcal{R}_j^u(t) = \sum_{i=1}^{U_1} \sum_{k=1}^{K} \frac{n_i^k(t) e_i^k(t)}{S \times U_2} B \log_2(1 + \frac{p_j^k(t) h_j^k(t)}{B \times N_0}) - \zeta_j^k(t)), \tag{3}$$

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 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.

The channel gain is assumed to be known with errors, the imperfection of channel estimation is modeled as follows

$$h_j^k(t) = \hat{h}_j^k(t) + \Delta h_j^k(t) \tag{4}$$

Where, $\Delta h_j^k(t)$ denotes the estimation error with a Guassian distribution of

$$\Delta h_j^k(t) \backsim \mathcal{N}(0, \boldsymbol{\phi}^{\boldsymbol{k}^2}), \tag{5}$$

C. Mean Delay

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

$$T^{\text{proc}} = T^{RU} + T^{DU} + T^{CU}, \tag{6}$$

Assume the packet arrival of URLLC UEs follows a Poisson process with arrival rate λ_j for the j^{th} UE. Therefore, the mean arrival data rate of the O-CU-UP layer is $\alpha^C = \sum_{j=1}^{U_2} \lambda_j$. 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 $\alpha = \alpha^C \approx \alpha^D$, 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

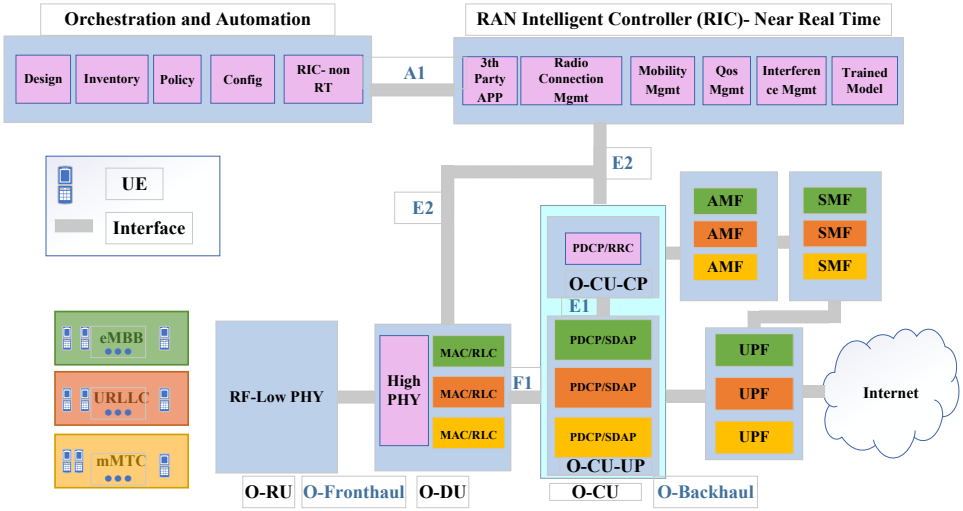

Fig. 1: Network sliced ORAN system

as M/M/1 queue, is formulated as follows, respectively [2]–[4],

$$T^{DU} = \frac{1}{\mu^d - \alpha/M^d}, \\ T^{CU} = \frac{1}{\mu^c - \alpha/M^c}, \tag{7}$$

where M^d,and M^c 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 $l \in \{d, c\}$.

T_j^{RU} is the mean transmission delay of the j^{th} UE of the URLLC service on the wireless link. The arrival data rate of wireless link for each UE j of URLLC service is λ_j As a result, we have $\sum_{j=1}^{U_2} \lambda_j = \alpha$. Moreover, The service time of transmission queue for UE j requesting URLLC service has an exponential distribution with mean $1/R_j^u$ 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

$$T_j^{RU} = \frac{1}{R_j^u - \lambda_j}. \tag{8}$$

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,

$$P_r\{T_j^{RU} \geq T_{RU}^{max}\} = e^{-(R_{tot}^u - \alpha)T_{RU}^{max}} \tag{9}$$

where, $R_{tot}^u = \sum_{j=1}^{U_2} R_j^u$ Also, we do not consider the reliability for O-CU and O-DU.

E. Problem Statement

The optimization problem is formulated as follow

$$\max_{\boldsymbol{E},\boldsymbol{N},\boldsymbol{M}} \sum_i R_i^e(t) - \eta \sum_j T^{\text{proc},u}(t) \tag{10a}$$

$$\text{subject to} \quad \mathcal{R}_i^e(t) \geq \mathcal{R}_{min}^e \quad \forall i \in U_1, \tag{10b}$$

$$Pr\{R_i^e(t) \leq R_{min}^e\} \leq \epsilon \tag{10c}$$

$$T_j^{\text{proc},u}(t) \leq T_{min}^u \quad \forall j \in U_2 \tag{10d}$$

$$Pr\{T_j^{\text{proc},u}(t) \geq T_{min}^u\} \leq \epsilon \quad \forall j \in U_2 \tag{10e}$$

where, (10c), supports the reliability of eMBB while puncturing the URLLC.

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.