TABLE I
VARIABLES AND OBJECTIVES OF DIFFERENT SLICES

Types of slices	Variables	Objectives of slices
Set definition	Δ (s) TTI slot length with set $\mathcal{T}$; - δ (s) minislot length with M the number of minislot per slot.	
High bandwidth slice	- Subchannel allocation ($\beta_{e,k}^{(t)}$); - downlink power ($p_{e,k}^{(d)}$)	- Total downlink sum rate
AI traffic-based slicing	- Time required for communication ($t_{n,m}$); - Time required for computation (f_n); - Communication power (p_n)	- Convergence rate; - Model accuracy
Ultra-reliable-based slicing	- SNR $\gamma^{(n)}$; - Target error probability $e^{(n)}$	- availability (α); - reliability(τ)

Simple case:

- Within one cell (BS)
- FL clients data iid and balanced. (only differ in computational capacity)

More advanced case I:

- Within one cell (BS)
- FL clients data niid and non balanced and also differ in computational capacity.

More advanced case II:

- Multi-cell (BSs)
- FL clients data niid and non balanced and also differ in computational capacity.

More advanced case III:

- Multi-cell (BSs)
- FL clients data niid and non balanced and also differ in computational capacity.
- Model-model FL training

In introduction, first more general as AI task/traffic in general, later specify e.g. FL.

I. System Model

A. Basic Settings

- Define Δ TTI slot length, $\mathcal{T}$ the slot set; $\delta = \Delta/M$ mini-slot length.
- $\mathcal{E}$ eMBB user service set.
- $\mathcal{U}$ URLLC user service set.
- $\mathcal{F}$ entire Federated Learning (FL) user service set and $\mathcal{S}_n \subset \mathcal{F}$ the training set for the communication round n, $\mathcal{F} \subset \mathcal{E} \cup \mathcal{U}$
- $\mathcal{K}$ frequency RB (subchannels).
- Traffic: Denote $a_e(t)$ the traffic generated at t TTI slot for UE $e \in \mathcal{E}$ and $a_u^{(t)}(m)$ the traffic generated at m-th minislot of the t-th TTI slot for $u \in \mathcal{U}$.
- *Rate per RB:*
  - the eMBB user downlink rate at RB k, $r_{e,k}^{(t)}$, is

$$r_{e,k}^{(t)} = B_k \log_2\left(1 + \gamma_{e,k}^{(t)}\right) \tag{1}$$

    where B_k is the bandwidth of RB k, and $\gamma_{e,k}^{(t)} = \frac{p_{e,k}^{d} h_{e,k}}{\sigma^2}$ the SNR (SINR for multiple BSs) of the channel BS-UE e at TTI slot t at RB k.
  - The FL broadcasting rate: denoted with index $s \in \mathcal{S}_n$: $r_{s,k}^{(t,dl)}$, where $\gamma_{s,k}^{(t,dl)}$ the SNR of downlink.
  - The FL update rate: denoted with index $s \in \mathcal{S}_n$: $r_{s,k}^{(t,ul)}$, where $\gamma_{s,k}^{(t,ul)}$ the SNR.
- *Total rate without puncturing:* Given $\beta_{e,k}^{(t)} \in \{0,1\}$ (resp. $\beta_{s,k}^{(t)} \in \{0,1\}$) the k RB allocation vector for $e \in \mathcal{E}$ (resp. $s \in \mathcal{S}_n$). Denote the

$$K_e^{(t)} = \sum_k \beta_{e,k}^{(t)}, \; K_s^{(t)} = \sum_k \beta_{s,k}^{(t)}, \tag{2}$$

  total number of RBs allocated to eMBB UE e (resp. s). It has to be satisfied that the total number of allocated RBs is smaller than K:

$$\sum_k \left[\sum_e \beta_{e,k}^{(t)} + \sum_{s\in\mathcal{S}_n} \beta_{s,k}^{(t)}\right] = \sum_e K_e^{(t)} + \sum_{s\in\mathcal{S}_n} K_s^{(t)} \le K \tag{3}$$

  With such notations, the achievable rate at TTI t without puncturing is:

$$r_e^{(t)} = \sum_k \beta_{e,k}^{(t)} r_{e,k}^{(t)} \tag{4}$$

- At each TTI slot t, $a_e(t)$ is the amount of traffic generated by the user $e \in \mathcal{E}$ and $r_e^{(t)}$ is the downlink data rate. ($a_e(t)$ i.i.d. over time) Let $q_e(t)$ represent the data packet backlog of the queue at BS for user $e \in \mathcal{E}$ at time t. For each TTI slot t, the $q_e(t)$ evolves according to following queue dynamics

$$q_e(t+1) = \max[q_e(t) - r_e^{(t)}, 0] + a_e(t) \tag{5}$$

- Further, for the queue of each user $e \in \mathcal{E}$ is called strongly stable if,

$$\lim_{T\to\infty} \sup \frac{1}{T} \sum_t \mathbb{E}[q_e(t)] < \infty \tag{6}$$

  For the system to be stable, all the individual queues of each user connected to the BS are strongly stable.

B. URLLC Puncturing

- The effect of URLLC puncturing on other services is purely based how many RBs it is needed (assume frequency-flat channels) and which other services user is being punctured. Therefore from the random variable $a_u^{(t)}(m)$ the traffic generated at m-th minislot of t-th TTI slot, to satisfy the URLLC requirement latency requirement, we assume that the URLLC traffic has to be transmitted within the minislot it is generated. (maybe should instead find the required number of blocklength,

then decide the minislot and frequency RB needed to satisfy the requirement.) Fix to numerology 1 so that the assumption of finishing the transmission within the next minislot holds.

- *Achievable rate of URLLC:* Assume uniform power allocation for subcarriers that ensures the same SNR, for $\omega \in \mathbb{N}$, the achievable rate of URLLC user that uses ω subcarriers for transmissions given the decoding error probability of $\varepsilon \in [0,1]$ is :

$$r_{u,\omega k}^{(t,m)} = \omega B\Big(\log_2(1+\gamma_{u,k}^{(t)}) - \sqrt{\frac{V_{u,k}^{(t)}}{\omega B\delta}}Q^{-1}(\varepsilon)\Big) \quad (7)$$

where $\gamma_{u,k}^{(t)} = \frac{p_{u,k}^d h_{u,k}}{\sigma^2}$ the SNR of URLLC user u; $V_{u,k}^{(t)} = (1 - \frac{1}{(1-\gamma_{u,k}^{(t)})^2})\log_2^2(e)$ the channel dispersion. The actual bits that are transmitted at minislot m of timeslot t is $\delta r_{u,\omega k}^{(t,m)}$.

Theorem 1. *Denoting $\omega_u^{(t)} \in \mathbb{N}$ the smallest feasible number of RB needed. For $a_u^{(t,m)} = 0$, clearly $\omega_u = 0$; otherwise, the following expression holds:*

$$\omega_u = \left\lceil \left(\frac{\sqrt{V}Q^{-1}(\varepsilon) + \sqrt{(\sqrt{V}Q^{-1}(\varepsilon))^2 + 4\log(1+\gamma_{u,k}^{(t)})a_u^{(t,m)}}}{2\sqrt{B\delta}\log(1+\gamma_{u,k}^{(t)})} \right)^2 \right\rceil \quad (8)$$

Proof. With given target decoding error probability ε, to satisfy the latency constraint, it should hold that the achievable rate depending on $\omega \in \mathbb{N}$ satisfies $r_{u,\omega k}^{(t,m)}\delta \geq a_u^{(t,m)}$. The goal is to find for all $a_u^{(t,m)} > 0$, $\omega_u = \min\{w \in \mathbb{N} \mid r_{u,\omega k}^{(t,m)}\delta \geq a_u^{(t,m)}\}$.
We note first that for $a_u^{(t,m)} = 0$, $\omega_u = 0$.
Now for $a_u^{(t,m)} > 0$, the condition becomes solving a second-order polynomial inequality w.r.t. $\sqrt{B\delta\omega}$:

$$\log(1+\gamma)B\delta\omega - \sqrt{V}Q^{-1}(\varepsilon)\sqrt{B\delta\omega} - a_u^{(t,m)} \geq 0 \quad (9)$$

The polynomial is negative when $\sqrt{B\delta\omega} = 0$, the solution of the problem has to be the positive root since $\sqrt{B\delta\omega} \geq 0$. As ω_u is the smallest integer, therefore the ceiling function.
Note that the expression only holds for $a_u^{(t,m)} > 0$, since for $a_u^{(t,m)} = 0$, the solution to the problem is the smaller root $\omega_u = 0$ instead of the larger one. □

- At the minislot m, the total number of RBs that need to be punctured is

$$K_{puncture}^{(m)} = \sum_{u\in\mathcal{U}} \omega_u. \quad (10)$$

- Given $K_{puncture}^{(m)}$ the total number of RBs that needed to be punctured at the minislot m (from URLLC buffer state), the already assigned RBs to current eMBB $K_e^{(t)}$ and FL downlink traffic $K_{FL,dl}^{(t)}$, we can define a weight vector for the puncturing strategy. The loss of rate for $e \in \mathcal{E}$ at the minislot m is

$$\ell_e^{(t,m)} = \frac{\zeta_e^{(t,m)}}{MK_e(t)} r_e(t), \quad (11)$$

where $\zeta_e^{(t,m)}, \zeta_{FL,dl}^{(t,m)} \in \{0,1,\ldots,K_e^{(t)}(\text{resp.}K_{FL,dl}^{(t)})\}$ denote the number of UE e (or s)'s RB that is punctured. And it should satisfy that:

$$\sum_e \zeta_e^{(t,m)} + \zeta_{FL,dl}^{(t,m)} = K_{puncture}^{(m)} \quad (12)$$

Note that the **feasibility constraint** is that $\sum_e K_e^{(t)} + K_{FL,dl}^{(t)} \geq K_{puncture}^{(m)}$ for each minislot m.

- Following the approximation approach in [1], we 'relax' the integer constraint of $\alpha_e^{(t,m)}$ and divide by $K_{puncture}^{(m)}$ to bring the values of $\alpha_e^{(t,m)} \in [0, K_e^{(t)}/K_{puncture}^{(m)}]$ to a unit simplex. The approximated loss of rate:

$$\ell_e^{(t,m)} = \frac{\alpha_e^{(t,m)} K_{puncture}^{(m)}}{MK_e(t)} r_e(t). \quad (13)$$

And the total rate loss over TTI t is:

$$\ell_e^{(t)} = \frac{\sum_{m=1}^M \alpha_e^{(t,m)} K_{puncture}^{(m)}}{MK_e(t)} r_e(t). \quad (14)$$

- The URLLC packet arrival at the minislots are unknown at the boundary of TTI slot. Therefore, the expected loss for $e \in \mathcal{E}$ at TTI slot t is:

$$\bar{\ell}_e^{(t)} = \frac{\alpha_e^{(t)}\,\mathbb{E}[K_{puncture}^{(m)}]}{K_e(t)} r_e(t). \quad (15)$$

Or

$$\bar{\ell}_e^{(t)} = \frac{r_e(t)}{K_e(t)} \sum_{k=1}^{K} \mathbb{P}([K_{puncture}^{(m)} = k)k\,\mathbb{E}[\alpha_e^{(t)} | K_{puncture}^{(m)} = k]. \quad (16)$$

but distribution of $\alpha_e^{(t)}$?

- *Expected Value of $K_{puncture}^{(tot)}$:*

$$\mathbb{E}[K_{puncture}^{(m)}] = \sum_u \mathbb{E}[\omega_u]. \quad (17)$$

With $a_u^{(t,m)} \sim Pois(\lambda_u)$, the expected value of $\mathbb{E}[\omega_u]$ can be estimated as:

$$\mathbb{E}[\omega_u] = \sum_{d\in\mathbb{N}\setminus\{0\}} (b_u + \sqrt{b_u^2 + c_u d})^2 e^{-\lambda_u} \frac{\lambda_u^d}{d!} \quad (18)$$

where $b_u = \frac{\sqrt{V}Q^{-1}(\varepsilon)}{2\sqrt{B\delta}\log(1+\gamma_{u,k}^{(t)})}$ and $c_u = \frac{1}{B\delta\log(1+\gamma_{u,k}^{(t)})}$. Because of lack of closed-form expression due to the presence of square root, approximation is needed. By using the inequality $(a+b)^2 \leq 2a^2+2b^2$ for every $a, b \in \mathbb{R}$, we obtain the following estimation:

$$\begin{aligned}\mathbb{E}[\omega_u] &\leq \sum_{d\in\mathbb{N}\setminus\{0\}} (2b_u^2 + 2(b_u^2 + c_u d))e^{-\lambda_u}\frac{\lambda_u^d}{d!} \\ &\leq 4b_u^2(1-e^{\lambda_u}) + 2c_u\lambda_u.\end{aligned} \quad (19)$$

The expected value of $K_{puncture}^{(tot)}$ is therefore upper bounded by:

$$\mathbb{E}[K_{puncture}^{(m)}] \leq \sum_{u\in\mathcal{U}} [4b_u^2(1-e^{\lambda_u}) + 2c_u\lambda_u] \quad (20)$$

- Actual puncturing: from $\{\alpha_e^{(t,m)}\}_e$ to actual punctuation $\zeta_e^{(t)}$. Unknown from [1]. Maybe weighted

random without replacement, or truncated multinomial distribution?

According to $a_u^{(t,m)}$ at the minislot m, the amount of RBs that is sufficient to satisfy URLLC constraint $K_{puncture}^{(m)}$ is computed. Then the sampling strategy with the weights/probability $\{\alpha_e^{(t)}\}_e$ to obtain $\zeta_e^{(t,m)}$.

- Using Eq. (2), the actual total eMBB rate of user e for the TTI slot t is:

$$r_e^{(t)} = \left(1 - \frac{\sum_m \zeta_e^{(t,m)}}{M K_e^{(t)}}\right) K_e^{(t)} r_{e,k}^{(t)} \tag{21}$$

The actual number of bits that can be transmitted at the time slot t is $r_e^{(t)}\Delta$.

We assume that the puncturing can only happen for the eMBB traffic, FL downlink part. (not FL uplink)

The expected eMBB rate with the allocation weight $\alpha_e^{(t)}$ is:

$$\overline{r}_e^{(t)} = \left(1 - \frac{\alpha_e^{(t)} \mathbb{E}[K_{puncture}^{(m)}]}{K_e^{(t)}}\right) K_e^{(t)} r_{e,k}^{(t)} \tag{22}$$

- The same goes for the downlink FL traffic. For t where FL is in the downlink phase, the rate loss can be written as:

$$\overline{r}_s^{(t)} = \left(1 - \frac{\alpha_{FL,dl}^{(t)} \mathbb{E}[K_{puncture}^{(m)}]}{K_{FL,dl}^{(t)}}\right) K_{FL,dl}^{(t)} r_{s,k}^{(t,downlink)} \tag{23}$$

and it actually satisfies that:

$$\sum_e \alpha_e^{(t)} + \alpha_{FL,dl}^{(t)} = 1 \tag{24}$$

- When FL is in the uplink phase, we assume that the allocated resources cannot be punctured.

[1] it can happen that the downlink and uplink happen at the same time... It is more UE-related. Denote a set of current downlink and uplink user, and define the previous rate and everything with the new notation

The URLLC reliability is defined as two parts (similar statement also exists in [2]): (i) decoding error probability; (ii) the needed RBs to puncture is greater than the RBs we can puncture. Denote $K' \in \mathbb{N}$ the RBs we are allowed to puncture. It should be satisfied that:

$$\mathbb{P}\left(K_{puncture}^{(m)} > K'\right) \leq \varepsilon_{puncture} \tag{25}$$

Theorem 2. *The constraint* (25) *can be satisfied when* $K' \geq K_{\min}$, *where* $K_{\min} = \frac{\sum_{u\in\mathcal{U}}[4b_u^2(1-e^{\lambda_u})+2c_u\lambda_u]}{\varepsilon_{puncture}}$. *Or* $\overline{K}'_{\min}$ *satisfies:*

$$g(\overline{K}'_{\min}, h^{-1}(\overline{K}'_{\min})) = \ln(\varepsilon_{puncture}), \tag{26}$$

where

$$g : (K, \mu) \mapsto -\mu K' - \sum_{u\in\mathcal{U}} \lambda_u + 4\mu \sum_{u\in\mathcal{U}} b_u^2 + \sum_{u\in\mathcal{U}} \lambda_u e^{2c_u\mu} \tag{27}$$

and

$$h : \mu \mapsto 4\sum_{u\in\mathcal{U}} b_u^2 + 2\sum_{u\in\mathcal{U}} \lambda_u c_u e^{2c_u\mu}. \tag{28}$$

Proof. By Markov's inequality, the probability term in (**??**) can be bounded as:

$$\begin{aligned}\mathbb{P}\left(\sum_u \omega_u \geq K'\right) &\leq \frac{\sum_u \mathbb{E}[\omega_u]}{K'} \\ &\leq \frac{\sum_{u\in\mathcal{U}}[4b_u^2(1-e^{\lambda_u})+2c_u\lambda_u]}{K'}\end{aligned} \tag{29}$$

Therefore, it is sufficient that:

$$K' \geq \frac{\sum_{u\in\mathcal{U}}[4b_u^2(1-e^{\lambda_u})+2c_u\lambda_u]}{\varepsilon_{puncture}} \triangleq K_{\min} \tag{30}$$

Tighter bounds is needed if $\varepsilon_{puncture}$ is very small (URLLC). By Chernoff's bound, we have for $\mu > 0$:

$$\mathbb{P}\left(\sum_u \omega_u \geq K'\right) \leq e^{-\mu K'} \prod_{u\in\mathcal{U}} \mathbb{E}[e^{\mu(b_u + \sqrt{b_u^2 + c_u a_u^{(t,m)}})^2}]. \tag{31}$$

The expected value term can be bounded as follows:

$$\mathbb{E}[e^{\mu(b_u+\sqrt{b_u^2+c_u a_u^{(t,m)}})^2}] \tag{32}$$

$$= \sum_{d\in\mathbb{N}\setminus\{0\}} e^{\mu(b_u+\sqrt{b_u^2+c_u d})^2} e^{-\lambda_u} \frac{\lambda_u^d}{d!} + e^{-\lambda_u} \tag{33}$$

$$\leq e^{-\lambda_u} + e^{-\lambda_u} \sum_{d\in\mathbb{N}\setminus\{0\}} e^{\mu(2b_u^2+2b_u^2+2c_u d)} \frac{\lambda_u^d}{d!} \tag{34}$$

$$\leq e^{-\lambda_u} + e^{-\lambda_u+4\mu b_u^2} \sum_{d\in\mathbb{N}\setminus\{0\}} \frac{(\lambda_u e^{2\mu c_u})^d}{d!} \tag{35}$$

$$\leq e^{-\lambda_u} + e^{-\lambda_u+4\mu b_u^2}(\exp(\lambda_u e^{2\mu c_u}) - 1) \tag{36}$$

$$\leq e^{-\lambda_u} + \exp(-\lambda_u + 4\mu b_u^2 + \lambda_u e^{2\mu c_u}) - e^{-\lambda_u+4\mu b_u^2} \tag{37}$$

$$\leq \exp(-\lambda_u + 4\mu b_u^2 + \lambda_u e^{2\mu c_u}) \tag{38}$$

The Chernoff's bound becomes:

$$\mathbb{P}\left(\sum_u \omega_u \geq K'\right) \leq e^{-\mu K'} \exp(\sum_{u\in\mathcal{U}} -\lambda_u + 4\mu b_u^2 + \lambda_u e^{2\mu c_u}). \tag{39}$$

$$\leq \exp(-\mu K' - \sum_{u\in\mathcal{U}} \lambda_u + 4\mu \sum_{u\in\mathcal{U}} b_u^2 + \sum_{u\in\mathcal{U}} \lambda_u e^{2\mu c_u}) \tag{40}$$

Since the inequality holds for all $\mu > 0$, we need to find the minimum w.r.t. μ. Define the function g as in [XX]. Given K', the function $\mu \mapsto g(K', \mu)$ is clearly convex. The minimum takes place either at $\frac{\partial g}{\partial \mu}(K', \mu^*) = 0$, i.e.,

$$K' = 4\sum_{u\in\mathcal{U}} b_u^2 + 2\sum_{u\in\mathcal{U}} \lambda_u c_u e^{2c_u\mu^*} \tag{41}$$

or if such point does not exists, i.e.

$$K' < 4\sum_{u\in\mathcal{U}} b_u^2 + 2\sum_{u\in\mathcal{U}} \lambda_u c_u \tag{42}$$

then $\mu^* = 0$ and no tight bound can be derived.

Due to the transcendent nature of Equation (41), no closed-form solution of μ^* can be found. But by the strict monotonicity of the right hand side of eq. (41), there is a bijective mapping from μ^* to K', denoted as $h : \mu \mapsto K'$ the expression in eq. (41).

To satisfy constraint (25), it is sufficient to have

$$\exp(-\mu^* K' - \sum_{u\in\mathcal{U}} \lambda_u + 4\mu^* \sum_{u\in\mathcal{U}} b_u^2 + \sum_{u\in\mathcal{U}} \lambda_u e^{2\mu^* c_u}) \leq \varepsilon_{puncture} \tag{43}$$

Lemma 1. *The function* $\Phi : K' \mapsto g(K', h^{-1}(K'))$ *is continuous and strictly decreasing.*

[2] Definition of g needs to be corrected.

By Lemma 1, there exists $\overline{K}'_{\min} > 0$ (actual $K'_{\min} = \lceil \overline{K}'_{\min} \rceil$) the minimum K' such that the URLLC reliability constraint is satisfied, i.e. for all $K' \geq \overline{K}'_{\min}$

$$\Phi(K') \leq \ln(\varepsilon_{puncture}), \tag{44}$$

with $\Phi(\overline{K}'_{\min}) = \ln(\varepsilon_{puncture})$. By the monotony of the function, we can efficiently find $\overline{K}'_{\min}$ with a bisection algorithm.

Proof of Lemma 1. Note that the expression of h^{-1} is unknown. Since h is continuously differentiable and strictly increasing, the inverse function theorem applies. Its derivative is calculated as follows:

$$\frac{d\mu}{dK'} = \frac{dh^{-1}(K')}{dK'} = \frac{1}{h'(h^{-1}(K'))}. \tag{45}$$

where we have $h'(\mu) = 4\sum_u \lambda_u c_u^2 e^{2c_u\mu}$ for $\mu \geq 0$. We obtain

$$\Phi'(K') = -\frac{K'}{h'(h^{-1}(K'))} - h^{-1}(K') + \frac{4\sum_u b_u^2}{h'(h^{-1}(K'))} + \frac{2\sum_u \lambda_u c_u e^{2c_u h^{-1}(K')}}{h'(h^{-1}(K'))} \tag{46}$$

By replacing the value of K' with the expression eq. (41), we obtain

$$\Phi'(K') = -h^{-1}(K') = -\mu^*(K') < 0 \tag{47}$$

□

□

C. Federated Learning Performance Characterization

We are interested in the communication parts of federated learning. The local training appears as a 'waiting time' before the information feedback, or for the FL uplink update, as a packet arrival time that we can control.

Downlink Phase: The communication round n starts at the TTI time step t_n, where the BS starts broadcasting the current model updates to all selected UEs $\mathcal{S}_n$. Let $K_s^{(t,dl)} \in [K]$ be the total allocated RBs allocated for the downlink broadcasting. After a UE $s \in \mathcal{S}_n$ has well received the whole packet, i.e., in average(To be sure)

$$\overline{T}_s^{(n,dl)} = \frac{D}{K_s^{(t,dl)} \Delta \mathbb{E}_{h_s}[r_{s,k}^{(t)}]} \text{ TTI slots.} \tag{48}$$

The actual finish TTI slot of the broadcasting of $s \in \mathcal{S}_n$ is:

$$T_s^{(n,dl)} = \min\left\{T \in \mathbb{N} \mid \Delta K_s^{(t,dl)} \sum_{t=t_n}^{T} r_{s,k}^{(t)} \geq D\right\} \tag{49}$$

The total downlink phase lasts until all the selected UEs $\mathcal{S}_n$ finished.

$$T^{(n,dl)} = \max_{s\in\mathcal{S}_n} T_s^{(n,dl)}. \tag{50}$$

We denote the subset of $\mathcal{S}_n$ that is doing downlink transmission at the TTI slot t as $\mathcal{S}_n^{(dl)}(t)$.

[3] Should $K_s^{t,dl}$ be dependent on $\mathcal{S}_n^{(dl)}(t)$? (should be)

Local Training Update: Each UE starts the local training after receiving the whole model (D bits) during a time denoted by $\tau_{s,n}^{(comp)}(f_{s,n})$ that is tunable by the computational capacity $f_{s,n}$.

Model uplink update: After the local update, each user 'asynchronously' requests the BS to transmit the updated model back to BS. We denote such UE at the TTI slot t as $\mathcal{S}_n^{(ul)}(t)$. Given $\mathcal{S}_n^{(ul)}(t)$, $K_s(t)$ and $K_e(t)$ need to be reallocated accordingly to:

1) Have enough RBs for FL to satisfy the long-term latency constraint.
2) Save enough RBs for eventual URLLC puncturing.
3) The eMBB should also not be too much penalized.

Denote $n \in \mathcal{N}$ communication round, $k \in \mathcal{K}$ end-devices, $m \in \mathcal{M}$ BSs. For communication round n, the total spent delay can be expressed as:

$$\tau_n^{(tot)} = \max_{s\in\mathcal{S}_s}\{\tau_{s,n}^{(comm,d)}(p^{(d)}) + \tau_{s,n}^{(comp)}(f_{s,n}) + \tau_{s,n}^{(comm,u)}(r_s^{(t)})\}. \tag{51}$$

The total spent energy in communication round n:

$$E_n^{(tot)} = \sum_{s\in\mathcal{S}_s} E_{s,n}^{(comm,d)}(p^{(d)}) + E_{s,n}^{(comp)}(f_{s,n}) + E_{s,n}^{(comm,u)}(r_s^{(t)}, p_{s,n}^{(t)}). \tag{52}$$

For total spent delay and energy over all FL communication round:

$$\tau^{(tot)} = \sum_n \tau_n^{(tot)}, \tag{53}$$

$$E^{(tot)} = \sum_n E_n^{(tot)} \tag{54}$$

- Time required for communication:

$$\tau_{s,n}^{(comm,d)} = \frac{d}{r_{s,n}^{(d)}}. \tag{55}$$

- Energy for communication:

$$E_{s,n}^{(comm,d)} = \tau_{s,n}^{(comm,d)} p_{s,n}^{(t)}. \tag{56}$$

- Time required for computation: with $f_{s,n} \in (f_{s,\min}, f_{s,\max})$ (potential variable),

$$\tau_{s,n}^{(\text{comp})} = \alpha \frac{D_s}{f_{s,n}}, \tag{57}$$

where $\alpha > 0$ some constant, D_s the local dataset size of UE s (with the same number of epochs, the actual number iteration is proportional to local dataset size), f_s the computation capacity user s.
- Energy of computation:

$$E_{s,n}^{(\text{comp})}(f_{s,n}) = \kappa\alpha D_s f_{s,n}^2, \tag{58}$$

- Since n and t are different time scales, we need to address this carefully. Denote T_n the TTI step that the communication round n starts, i.e., also the TTI step where the previous communication round $n-1$ ends. Given T_n, the duration of this communication round (i.e. $T_{n+1} - T_n$) is hard to predict, since it depends on the channel fading, packet

General time-average problem,

$$\min_{\substack{\{\mathcal{S}_n(t),f_{s,n}(t),p^{(u)}_{s,n}(t)\}_{s,n,t}\\ \{K^{(t)}_e,K^{(t)}_{FL,dl},K^{(t)}_{s,ul},\alpha^{(t)}_e,\alpha^{(t)}_{FL,dl}\}_{t,e,s\in\mathcal{S}_n}}} \quad \Big\{\sum_{n=1}^{N}\max_{s\in\mathcal{S}_n}\{\tau^{(tot)}_{s,n}\}, \quad -\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\sum_{e\in\mathcal{E}}\mathbb{E}_{\boldsymbol{a}_u}[r^{(t)}_{e,k}]\Big\} \tag{59a}$$

$$\text{s.t.} \quad (\forall e\in\mathcal{E}) \quad \mathbb{P}(q_e(t)\geq q_0)\leq\varepsilon_q \tag{59b}$$

$$(\forall s\in\mathcal{F}) \quad E^{(tot)}_s \triangleq \sum_n E^{(tot)}_{s,n} \leq E_{\text{budget},s}, \tag{59c}$$

$$(\forall t), \ \mathbb{P}\left(K^{(m)}_{puncture}\geq\sum_e K^{(t)}_e + K^{(t)}_{FL,dl}\mathbb{1}_{\mathcal{S}^{(dl)}_n(t)\neq\emptyset}\right)\leq\varepsilon_{puncture} \tag{59d}$$

$$(\forall n,t) \quad \sum_e K^{(t)}_e + K^{(t)}_{FL,dl}\mathbb{1}_{\mathcal{S}^{(dl)}_n(t)\neq\emptyset} + \sum_{s\in\mathcal{S}^{(ul)}_n(t)} K^{(t)}_{s,ul}\leq K \tag{59e}$$

$$(\forall t) \quad \sum_e \alpha^{(t)}_e + \alpha^{(t)}_{FL,dl} = 1 \tag{59f}$$

$$(\forall e\in\mathcal{E},\ \forall t) \quad 0\leq\alpha^{(t)}_e\leq\frac{K^{(t)}_e}{\mathbb{E}[K^{(m)}_{puncture}]} \tag{59g}$$

$$(\forall t) \quad 0\leq\alpha^{(t)}_{FL,dl}\leq\frac{K^{(t)}_{FL,dl}}{\mathbb{E}[K^{(m)}_{puncture}]} \tag{59h}$$

$$(\forall n) \quad |\mathcal{S}_n(t)|\geq C_n \tag{59i}$$

$$\Big(\forall n,\ \forall s\in\mathcal{S}_n\backslash(\mathcal{S}^{(dl)}_n(t)\cup\mathcal{S}^{(ul)}_n(t))\Big) \quad f_{s,n}(t)\in[f_{s,min},f_{s,max}] \tag{59j}$$

$$(\forall n,\ \forall s\in\mathcal{S}^{(ul)}_n(t)) \quad p^{(ul)}_{s,n}(t)\in[0,P_{\max}] \tag{59k}$$

$$\sum_{n=1}^{N}\max_{s\in\mathcal{S}_n}\{\tau^{(tot)}_{s,n}\}\leq T_{FL,budget} \tag{59l}$$

$$\min_{\substack{\{\mathcal{S}_n(t),f_{s,n}(t),p^{(u)}_{s,n}(t)\}_{s,n,t}\\ \beta^{(t)}_{s,k},\beta^{(t)}_{e,k},\overline{\alpha}^{(t)}_e,\overline{\alpha}^{(t)}_{s,n}\}_{m,k,t}}} \quad \sum_{n=1}^{N}\max_{s\in\mathcal{S}_n}\{\tau^{(tot)}_{s,n}\} - \lambda\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\sum_{e\in\mathcal{E}}\mathbb{E}_{\boldsymbol{a}_u}[r^{(t)}_{e,k}] \tag{60a}$$

$$\text{s.t.} \quad (\forall e\in\mathcal{E}) \quad \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}q_e(t)\leq\varepsilon_q q_0 \tag{60b}$$

$$(\forall s\in\mathcal{F}) \quad E^{(tot)}_s \triangleq \sum_{n=1}^{N} E^{(tot)}_{s,n} \leq E_{\text{budget},s}, \tag{60c}$$

$$(\forall t), \ \sum_e K^{(t)}_e + K^{(t)}_{FL,dl}\mathbb{1}_{\mathcal{S}^{(dl)}_n(t)\neq\emptyset}\geq\overline{K}_{\min}(t) \tag{60d}$$

$$(\forall n,t) \quad \sum_{e\in\mathcal{E}} K^{(t)}_e + K^{(t)}_{FL,dl}\mathbb{1}_{\mathcal{S}^{(dl)}_n(t)\neq\emptyset} + \sum_{s\in\mathcal{S}^{(ul)}_n(t)} K^{(t)}_{s,ul}\leq K \tag{60e}$$

$$(\forall t) \quad \sum_{e\in\mathcal{E}} \alpha^{(t)}_e + \alpha^{(t)}_{FL,dl} = 1 \tag{60f}$$

$$(\forall e\in\mathcal{E},\ \forall t) \quad 0\leq\alpha^{(t)}_e\leq\frac{K^{(t)}_e}{\mathbb{E}[K^{(m)}_{puncture}]} \tag{60g}$$

$$(\forall t) \quad 0\leq\alpha^{(t)}_{FL,dl}\leq\frac{K^{(t)}_{FL,dl}}{\mathbb{E}[K^{(m)}_{puncture}]} \tag{60h}$$

$$(\forall n) \quad |\mathcal{S}_n(t)|\geq C_n \tag{60i}$$

$$\Big(\forall n,\ \forall s\in\mathcal{S}_n\backslash(\mathcal{S}^{(dl)}_n(t)\cup\mathcal{S}^{(ul)}_n(t))\Big) \quad f_{s,n}(t)\in[f_{s,min},f_{s,max}] \tag{60j}$$

$$(\forall n,\ \forall s\in\mathcal{S}^{(ul)}_n(t)) \quad p^{(ul)}_{s,n}(t)\in[0,P_{\max}] \tag{60k}$$

$$\sum_{n=1}^{N}\max_{s\in\mathcal{S}_n}\{\tau^{(tot)}_{s,n}\}\leq T_{FL,budget} \tag{60l}$$

How often should the problem be solved?

Add Variable Table?

Issues: Since FL is mostly uplink, if we puncture a FL uplink slot, FL might not have enough time to stop its transmission, leading to superposition/interference

II. Problem Formulation

Allocation problem without FL: Without FL, the allocation problem consists of

1) how many RBs to assign for each eMBB user
2) When there is URLLC traffic needed, which RB to puncture

Allocation problem with FL: With FL, the allocation problem consists of

1) how to balance RBs between eMBB and FL?
2) how to assign RB within each slice for eMBB and FL?
3) When there is URLLC traffic needed, which RB to puncture? eMBB or FL user? Which user?

III. Solutions

Multi-objective $->$ introduce $\lambda \geq 0$ to balance the two objectives to obtain a Pareto-optimal point.

It is a problem with different time scales, w.r.t. n and t. Each communication round n may contain variable numbers of time steps t. (renewal system)

Denoting $q_s^{(E)}(n)$ the virtual energy queue of $s \in \mathcal{F}$:

$$q_s^{(E)}(n+1) = (q_s^{(E)}(n) + E_{s,n}^{(tot)} - E_{\text{budget},s})^+ \tag{61}$$

Denote $q_s^{(T)}(n)$ the virtual time delay queue of $s \in \mathcal{F}$:

$$q^{(T)}(n+1) = (q^{(T)}(n) + \tau_n^{(tot)} - \frac{T_{FL,budget}}{N})^+ \tag{62}$$

For FL uplink traffic, the number of total RBs used should be very limited, since not only it affects the eMBB traffic, it may also impact the feasibility of the URLLC constraint. If only one RB is available, by first come first serve (FCFS), the makespan $T_{|\mathcal{S}_n|}$ of the process can be calculated as:

$$T_1 = \tau_1^{(dl)} + \tau_1^{(comp)} + \tau_1^{(ul)} \tag{63}$$
$$\dots \tag{64}$$
$$T_n = \max\{T_{n-1},\ \tau_n^{(dl)} + \tau_n^{(comp)}\} + \tau_n^{(ul)} \tag{65}$$
$$\dots, \tag{66}$$

assuming that all clients were ordered like $\tau_1^{(dl)} + \tau_1^{(comp)} < \tau_2^{(dl)} + \tau_2^{(comp)} < \ldots < \tau_{|\mathcal{S}_n|}^{(dl)} + \tau_{|\mathcal{S}_n|}^{(comp)}$. The makespan can be upper bounded by

$$T_{|\mathcal{S}_n|} \leq \max_{s\in\mathcal{S}_n}\{\tau_s^{(dl)} + \tau_s^{(comp)}\} + \sum_{s\in\mathcal{S}_n} \tau_s^{(ul)} \tag{67}$$

We solve the following per-round n problem:

$$\min_{\substack{\mathcal{S}_n, \overline{K}_e \\ \overline{K}^{(FL,dl)}, \overline{K}^{(FL,ul)},}} V\left[\max_{s\in\mathcal{S}_n}\{\frac{\mathbb{E}[\tau_s^{(dl)}]}{\overline{K}^{(FL,dl)}} + \tau_s^{(comp)}\} + \frac{\sum_{s\in\mathcal{S}_n}\mathbb{E}[\tau_s^{(ul)}]}{\overline{K}^{(FL,ul)}} - \lambda\sum_{e\in\mathcal{E}}\overline{K_e}\,\mathbb{E}_{\boldsymbol{\alpha},h_e}[r_{e,k}^{(t)}]\right\}\Bigg] + \sum_{s\in\mathcal{S}_n} q_s^{(E)}(n)E_{s,n}^{(tot)} \tag{68a}$$

$$\text{s.t.} \tag{68b}$$
$$|\mathcal{S}_n(t)| \geq C_n \tag{68c}$$
$$\overline{K_e}, \overline{K}^{(FL,dl)}, \overline{K}^{(FL,ul)} \geq 0 \tag{68d}$$
$$\sum_{e\in\mathcal{E}}\overline{K_e} + \overline{K}^{(FL,dl)} \geq \overline{K}'_{\min} \tag{68e}$$
$$\sum_{e\in\mathcal{E}}\overline{K_e} + \overline{K}^{(FL,dl)} + \overline{K}^{(FL,ul)} \leq K \tag{68f}$$

Need also to include puncturing effect already in to get a "target performance"?

If we can choose $K_e(t) = \overline{K}_e$, $K^{(FL,dl)}(t) = \overline{K}^{(FL,dl)}$, and $K_s^{(FL,ul)}(t) = \overline{K}_s^{(FL,ul)}$ for all t, then

Theorem 3. *For $t = 1, \ldots, T$, with a certain (stationary) random variable X and it has a new i.i.d. realization for every time step t, X_t. We have found the optimizer*

$$x^* = \arg\min_x \mathbb{E}_X[f_X(x)]$$

for some function f that has its parameter dependent on the random variable X.

This value $\mathbb{E}_X[f_X(x^*)]$ is what I desired:

$$\frac{1}{T}\sum_{t=1}^{T} f_{X_t}(x^*) \stackrel{!}{=} \mathbb{E}_X[f_X(x^*)]$$

However, for the "real-time" problem, several constraints are present. First, our x is an integer value when applied in practice, so cannot take x^* exactly; secondly, depending on the realization X_t, x^* may be not feasible for the "real-time" problem.

To compensate for these two reasons, assuming the actual action taken at time t, $\overline{x}_t$, we propose a performance gap queue $Q(t)$ s.t.:

$$Q(t+1) = Q(t) + f_{X_{t+1}}(x^*) - f_{X_{t+1}}(\overline{x}_{t+1})$$

as a virtual queue to optimize together with other objectives.

Optimization problem for each TTI slot t. The most critical part is the FL uplink transmission. Difficulties due to the different packet arrival time of UE s (downlink communication time + computation time), and different uplink transmission time. Depending on how many parallel queues (RBs) we can have in total for FL uplink, also depends on the scheduling algorithms, the ending time of this communication round (makespan) is hard to capture.

At each TTI slot t, the following problem is solved:

A. Without controlling client selection

B. Considering Client selection

IV. Thoughts

FL transmission steps:

1) Broadcasting: Assume occupy a RB and that cannot be punctured. The latency needed depends on the worst channel clients in $\mathcal{S}_n$.
2) After local training of τ_k for $k \in \mathcal{S}_n$, the UE k sends a sending request to BS asking for uplink transmission in the next TTI slot.
3) At the boundary of this TTI slot, a new UE is added in RB allocation problem. Since only downlink transmissions are considered for eMBB and URLLC, even if the UE requesting RB is also an other slice UE, as long as the RB is different the transmission is not impacted.

Possible FL uplink scheduling strategy:

- Sequential scheduling [3]

Due to the tight URLLC latency constraint, real-time sophisticated slot allocation for URLLC is impossible. We designed a probabilistic assignment for puncturing slot allocation that are solved at the boundary of each TTI slot. Since RBs themselves are indifferent for URLLC clients, only the other services clients that are using the RBs are relevant. We propose the allocation probability at each communication round n, $\alpha_e, \alpha_{s,n} \in [0, 1]$ with

$$\sum_{e \in \mathcal{E}} \alpha_e + \sum_{s \in \mathcal{S}_n} \alpha_s = 1 \tag{69}$$

A. System Analysis

Given all maximum power, frequency flat channel. Given a maximum available resource block of $K_{\mathcal{F}} \geq 1$. What is the latency of FL transmission?

Denote $r_s^{(t)}$ the maximum achievable rate of one RB. Consider first-come first-serve (FCFS) policy for FL client scheduling.

Denoting $r_s^{(downlink)}$, the minimum time required for each UE to update their local update is :

$$\tau_s = \frac{d}{r_s^{(downlink)}} + \tau_s^{(comp)}(f_{s,max}) \tag{70}$$

Maybe a threshold on the estimated downlink and computation time, so that the stragglers are forgotten.

REFERENCES

[1] M. Alsenwi, N. H. Tran, M. Bennis, A. Kumar Bairagi, and C. S. Hong, "eMBB-URLLC resource slicing: A risk-sensitive approach," *IEEE Communications Letters*, vol. 23, no. 4, pp. 740–743, 2019.

[2] R. Kassab, O. Simeone, and P. Popovski, "Coexistence of URLLC and eMBB services in the C-RAN uplink: An information-theoretic study," in *IEEE Global Communications Conference (GLOBECOM)*, 2018, pp. 1–6.

[3] K. Guo, Z. Chen, H. H. Yang, and T. Q. S. Quek, "Dynamic scheduling for heterogeneous federated learning in private 5G edge networks," *IEEE J. Sel. Topics Sig. Process.*, vol. 16, no. 1, pp. 26–40, 2022.