Per O-DU Slice & Mobility Management

We model *slice management and mobility management per O-DU*. In particular, for each O-DU, the controller solves a resource allocation problem that (i) determines which O-RUs are connected to each client (the slice connectivity) and (ii) sets, for every active O-RU–client link, the *per–slice power cap* that the client is allowed to use when uploading its FL model. The problem is re–solved at every near-RT control interval (i.e., every 400 ms). Below, we first describe our problem formulation and then its solution design.

Problem formulation of near-RT controller. Let $\mathcal{B}$ be the set of O-RUs and $\mathcal{U}_s$ the set of clients that execute FL-service $s \in \mathcal{S}$ (here, $\mathcal{S}$ denotes the set of all FL-services in the system). We denote each O-RU by index $b \in \mathcal{B}$ and each client by index $u \in \mathcal{U}_s$. For each pair (b, u), let $g_{b,u}$ denote the channel gain, $B_{b,u}$ denote the allocated bandwidth, $P_{b,u}$ denote the allocated transmit power, and $R_{b,u}$ denote the data rate. Also, let $N_0 > 0$ denote the noise power spectral density. Each O-RU b is presumed to have a bandwidth budget $\overline{B}_b$. Also, each client u is assumed to have a total transmit power budget $\overline{P}_u$ (W), an FL model with the size of n_u^{Serv} (bits), which is determined by its executed FL-service. For each FL service s, we consider a deadline $T_s^{\max}$ per global round (reflecting its FL-service requirement), by which all clients that execute FL-service s must transfer their models back to the server in uplink.[1] clients who fail to meet this deadline are considered unsuccessful, and their local models are excluded from model aggregation in that round. Furthermore, we assign an importance weight $w_s^{\mathsf{Serv}} \geq 0$ to each FL-service s, which reflects the relative priority/importance of the clients of that FL-service in utilizing the network resources.

We model near-RT controller as the following optimization problem:

$$\boldsymbol{P_1}: \max_{\substack{B_{b,u}\geq 0,\, P_{b,u}\geq 0,\, R_{b,u}\geq 0\\ 0\leq z_u\leq 1}} \quad \underbrace{\sum_{s\in\mathcal{S}} w_s^{\mathsf{Serv}} \sum_{u\in\mathcal{U}_s} z_u}_{\text{FL-Service Requirement Satisfaction}} \quad - \quad \underbrace{\lambda_E \sum_{b\in\mathcal{B}} \sum_{s\in\mathcal{S}} \sum_{u\in\mathcal{U}_s} P_{b,u}}_{\text{Power Consumption}} \tag{1}$$

$$\text{(O-RU bandwidth)} \quad \sum_{s\in\mathcal{S}} \sum_{u\in\mathcal{U}_s} B_{b,u} \leq \overline{B}_b, \qquad \forall\, b \in \mathcal{B}, \tag{2}$$

$$\text{(client transmit power)} \quad \sum_{b\in\mathcal{B}} P_{b,u} \leq \overline{P}_u, \qquad \forall\, s \in \mathcal{S}, \forall\, u \in \mathcal{U}_s, \tag{3}$$

$$\text{(Data rate)} \quad R_{b,u} \leq B_{b,u} \ln\Bigl(1 + \frac{g_{b,u}}{N_0} \frac{P_{b,u}}{B_{b,u}}\Bigr), \quad \forall\, b \in \mathcal{B}, \forall\, s \in \mathcal{S}, \forall\, u \in \mathcal{U}_s, \tag{4}$$

$$\text{(FL-Service Metric/Latency)}\; z_u \leq \frac{\sum_{b\in\mathcal{B}} R_{b,u}}{\frac{n_u^{\mathsf{Serv}}}{T_s^{\max}} \ln 2} \Rightarrow \sum_{b\in\mathcal{B}} R_{b,u} \geq \frac{n_u^{\mathsf{Serv}}}{T_s^{\max}} \ln 2 \;\; z_u, \quad \forall\, s \in \mathcal{S}, \forall\, u \in \mathcal{U}_s, \tag{5}$$

where $\ln 2$ in (5) converts bits to nats, making it consistent with the data rate computation in (4), which is in nats/s due to the use of ln(.) function (this unit is used as it makes the problem easily solvable in MOSEK solver). In the above problem, constraint (2) ensures the adherence to the total bandwidth of each O-RU while (3) ensures that multi–path transmissions across multiple O-RUs

[1] We omit the local computation time by presuming uniform computations across the clients, i.e., each FL-service asks for a unified clock CPU speed allocation across its recruited clients that leads to a uniform computation time that can be included in $T_s^{\max}$.

cannot exceed the total transmit power budget of each client. Also, constraint (4) determines the data rate bound and (5) enforces that each client uploads its model within the global-round deadline $T_s^{\max}$, where the degree to which the deadline is met is captured via $z_u \in [0,1]$. In particular, since wireless resources (i.e., bandwidth and transmit power) are scarce, some clients may be unable to meet their FL-services' deadlines. We subsequently relax the hard requirement on deadline satisfaction by scaling the per-client latency (minimum-rate) constraint in (5) with a decision variable $z_u \in [0,1]$, which acts as a continuous admission indicator ($z_u = 1$ means the deadline is met). Note that higher values of z_u reward the objective function in (1), which makes constraints (5) a tight constraint.

In EFL, our goal is to maximize the number (or weighted number, obtained based on the FL-service importance) of clients that can return their models to the server on time. To capture this, the aim of the optimization problem $\boldsymbol{P_1}$ is to select optimal per-link bandwidth $B_{b,u} \geq 0$, power $P_{b,u} \geq 0$, rate $R_{b,u} \geq 0$ (nats/s), and $z_u \in [0,1]$ to maximize $\sum_{s\in\mathcal{S}} \sum_{u\in\mathcal{U}_s} w_s^{\mathsf{Serv}} z_u$ (i.e., the first term in the objective function), while minimizing the power consumption overhead on the FL clients (i.e., the second term in the objective function).

Conic implementation. To solve $\boldsymbol{P_1}$, we note that constraint (4) is not in a standard form: a conic solver (e.g., MOSEK, which we have used to solve our problem) requires a cone-representable form. To address this, constraint (4) can be imposed via the following equivalent exponential cone:

$$\big(R_{b,u},\, B_{b,u},\, B_{b,u} + \beta_{b,u} P_{b,u}\big) \;\in\; \mathcal{K}_{\exp}, \qquad \beta_{b,u} = \frac{g_{b,u}}{N_0}, \quad \forall (b,u), \tag{6}$$

with $\mathcal{K}_{\exp} = \{(x,y,z) : y > 0,\; y e^{x/y} \leq z\} \cup \{(x,0,z) : x \leq 0,\; z \geq 0\}$. The program (1)–(5) with (6) is a exponential–cone program and is solved per O-DU by a conic solver (e.g., MOSEK).

client connectivity and per–slice power caps. Let $(B^\star_{b,u}, P^\star_{b,u}, R^\star_{b,u}, z^\star_u)$ be the optimal solution to the optimization problem $\boldsymbol{P_1}$ obtained as detailed above. We create an FL RAN slice to specify the set of FL clients connected to each O-RU and to define the maximum transmit power allowed for each client within that slice. Specifically, the *connection* between O-RU b and the client u will be established if the allocated power exceeds a small threshold:

$$(b,u) \text{ is connected} \iff P^\star_{b,u} \;\geq\; \tau_P = 0.01 \text{ W}. \tag{7}$$

Furthermore, for every connected link (b,u), $P^\star_{b,u}$ denotes the slice-level maximum transmit power of client u, which is allowed to be used within that slice to transmit a portion of its local model. This power cap will subsequently be used by the MAC scheduler to enforce an upper bound on the client's transmit power when sending portions of its local model to the server via O-RU b, as discussed below.

Per O-RU MAC scheduling. The MAC scheduling problem follows the same structure and conic encoding as $\boldsymbol{P_1}$, but it differs in three key ways: (i) it is solved *per O-RU* and only over the clients currently attached to that O-RU (determined by near-RT controller); (ii) the *client power on each link* is upper bounded by the *slice–level power cap* $P^\star_{b,u}$ computed by the near–RT controller, i.e., $0 \leq P_{b,u} \leq P^\star_{b,u}$; and (iii) the latency (minimum–rate) constraint is enforced *per link* as $R_{b,u} \geq \frac{n_u^{\mathsf{Serv}}}{T_s^{\max}} \ln 2\, x_u$ (here, x_u is the admission variable, analogous to z_u used in the near–RT slice and mobility management stage).