Two-level Soft RAN Slicing for Customized Services in 5G-and-beyond Wireless Communications

Abstract—In this paper, a two-level soft slicing scheme is proposed for 5G-and-beyond radio access networks (RAN) to support ultra-reliable and low-latency communications (URLLC) and enhanced mobile broadband (eMBB) services with delay/reliability and throughput requirements, respectively. At the network-level, we first determine the amount of radio resources required for eMBB services, and analyze the delay violation probability for URLLC services. Then, an integer nonlinear program is formulated for the network-level resource pre-allocation. Since the formulated problem is NP-complete, a low-complexity heuristic algorithm is proposed to obtain near-optimal solutions. Given the pre-allocated resources at each gNodeB (gNB), a gNB-level resource scheduling scheme is designed to enable real-time resource sharing among URLLC services considering the reliability and delay requirements. Simulation results show that the proposed soft slicing scheme meets stringent quality-of-service (QoS) requirements for both URLLC and eMBB services, and achieves high resource utilization efficiency when compared with conventional hard resource slicing schemes.

Index Terms—RAN slicing, Dynamic scheduling, URLLC, eMBB, Industrial IoT.

I. INTRODUCTION

The 5G-and-beyond radio access networks (RAN) are foreseen to accommodate various emerging services [1], i.e., enhanced mobile broadband (eMBB) services and ultra-reliable and low-latency communication (URLLC) services. Particularly, eMBB services require high data rates, typical applications include multimedia, high definition video streaming and virtual reality (VR). URLLC services require less than 1 ms user plane latency and higher than 99.999% reliability in terms of packet transmission [2], which is crucial to many industrial Internet-of-things (IIoT) applications, such as factory automation, remote robotic control [3]. Besides, the scheduling interval for some IIoT URLLC services is in mini-slot level with a duration as short as 0.125 ms [4]. For the services with diverse traffic characteristics and quality-of-service (QoS) requirements [5], an enhanced resource management solution is required for the 5G-and-beyond RAN to ensure distinct QoS while achieving high resource utilization efficiency.

As one of the fundamental enablers to 5G and future networks, network slicing can dynamically form multiple virtual networks (slices) over a shared physical substrate, where each virtual slice maintains a proprietary combination of resources (e.g., communication, computing and caching resources) to guarantee the QoS requirements of supported services [6] [7]. For the RAN slicing, each virtual slice is assigned with a certain amount of radio resources, i.e., resource blocks (RBs), from a virtual RB pool shared among a group of gNodeBs (gNBs) [8] [9]. To realize QoS isolation among different virtual RAN slices and maximize the resource utilization at the same time, a network-wide RAN resource slicing scheme is required, which can dynamically adjust the amount of RBs allocated to each slice according to network conditions [10].

We aim at developing an efficient RAN slicing scheme to ensure the reliability and latency requirements of URLLC services, while guaranteeing minimum average throughput of eMBB services [11], [12]. Jointly accommodating URLLC and eMBB services in 5G-and-beyond RAN faces technical challenges: First, the stringent delay and reliability requirements of URLLC services, amongst the dynamic RB requests varying in a mini-slot level, are difficult to be guaranteed simultaneously. Second, the network-level resource slicing among gNBs can achieve the global optimum, but frequent executions of resource allocation process in each slice leads to large signaling overhead between gNBs and control entities (e.g., the centralized controller for radio resource slicing among gNBs based on software-defined networking (SDN) and network function virtualization (NFV) techniques). Therefore, it is challenging to balance the trade-off between global optimality for resource slicing and fast response to traffic variations. Third, instead of assigning a fixed amount of resources to each gNB or service, a slicing scheme that allows dynamic inter-slice resource sharing, i.e., soft-slicing, is desired to increase the multiplexing gain with guaranteed QoS [13]. Given the mini-slot level URLLC traffic dynamics, a soft slicing scheme with mini-slot level inter-gNB resource sharing is required.

In this paper, we propose an SDN/NFV enabled two-level (i.e., network-level and gNB-level) soft RAN slicing scheme to accommodate both URLLC and eMBB services over multiple gNBs, and to support dynamic inter-gNB RB sharing to exploit the resource multiplexing gain. In the network-level, we aim at minimizing the sum of RBs pre-allocated to both URLLC and eMBB services while guaranteeing their QoS requirements. Different from the “hard-slicing” schemes which allocate fixed RB set to each service, in this work, the RBs pre-allocated to each service can be accessed by other services opportunistically. Given the RBs pre-allocated to each service on each gNB, the fine-grained RB scheduling and sharing are realized in the gNB-level. The pre-allocated RBs are scheduled to URLLC/eMBB data transmission in each mini-slot/slot. Specifically, the gNB-level scheduling scheme executed on each gNB not only schedules pre-allocated RBs to different service requests for QoS-guarantee in real-time operations, but also enables collision-free inter-gNB RB sharing, i.e., each service under a specific gNB is allowed to temporarily access available RBs from other gNBs. The main contributions of this work are three-folded:

1) We propose a hierarchical soft RAN-slicing scheme, consisting of network-level RB pre-allocation and gNB-level RB scheduling. The network-level slicing reserves RBs to services for both QoS isolation and efficient RB
utilization. The gNB-level RB scheduling dynamically assigns RBs to URLLC and eMBB nodes and enables real-time RB sharing among gNBs;
2) For the network-level RB pre-allocation, the amount of RBs pre-allocated to each service is optimized considering QoS requirements and the probabilities of sharing RBs with other gNBs, which achieves both service isolation and resource multiplexing. For the gNB-level RB scheduling, a collision-free RB scheduling scheme is designed to guarantee stringent QoS requirements of URLLC services, while enabling mini-slot level inter-gNB resource sharing;
3) The network-level RB pre-allocation problem is formulated as an integer nonlinear program (INLP). The delay violation probability constraints for URLLC services and the throughput constraints for eMBB services are incorporated in the formulated problem through queuing theories and the random waypoint (RWP) mobility model, respectively. A low-complexity heuristic algorithm is proposed to address the NP completeness of the formulated problem.

The remainder of this paper is organized as follows: Related works are reviewed in Section II. The system model is described in Section III. The network-level resource pre-allocation problem is formulated in Section IV, followed by the heuristic algorithm. The gNB-level RB scheduling scheme is discussed in Section V. Simulations are presented in Section VI, and conclusions are drawn in Section VII.

II. RELATED WORK

RAN slicing is enabled by both NFV and SDN technologies. By using NFV, radio access and processing functions under gNBs are virtualized and centrally controlled by an NFV controller, such that the associated radio resources under gNBs are centrally pooled and managed [14]. The central controller is also SDN-enabled, which provides programmability on all gNBs to flexibly slice the pooled RAN resources among different services for more fine-grained QoS-oriented resource orchestration [15]. Depending on the application scenarios, various existing works formulate problems to jointly slice the communication, computing and caching resources of RAN [16], [17]. A fixed amount of resources are assigned to each slice for service isolation, which are referred to as “hard-slicing” schemes.

Based on “hard” RAN slicing schemes, some slices can be overloaded, while the resources of others are underutilized, given dynamic service requests. Therefore, soft-slicing schemes that allow the inter-slice resource sharing are proposed to increase the efficiency of resource utilization [18]-[20]. For communication resource slicing, a RAN slicing algorithm is proposed in [19] to leverage different resource abstraction types for a higher multiplexing gain. In [20], the earliest deadline first (EDF) scheduling mechanism is introduced to ensure QoS and improve resource utilization in real-time operation. However, the dynamics of RBs requests, which trigger the re-slicing at the beginning of each slicing window, are yet to be considered in determining the allocated resources for each slice. Moreover, the triggering of re-slicing due to changes of traffic arrival statistics cannot accommodate the small time scale traffic dynamics of certain services, e.g., URLLC, caused by traffic burstiness. Therefore, a hierarchical RAN slicing scheme that adapts to the traffic dynamics of multiple time scales from heterogeneous 5G-and-beyond services is required [13].

There exist many studies on resource allocation for URLLC services in IIoT scenarios, with different analytical traffic models and wireless transmission technologies [2]. Due to strict delay and reliability demands and mini-slot level traffic burstiness of URLLC services, allocating a fixed amount of communication resources can lead to low resource utilization or QoS dissatisfaction. Therefore, URLLC services should share resources dynamically with other services, such as eMBB, to exploit more resource multiplexing opportunities [10]. Customized RAN slicing with inter-slice sharing and mini-slot level resource scheduling is desired for accommodating a combination of URLLC and eMBB services.

III. SYSTEM MODEL

In this section, we present the system model for the soft RAN slicing problem. Table I lists the important symbols and notations used in the following sections.

A. RAN Model

Consider a downlink transmission system of an SDN/NFV enabled RAN, where a set of gNBs, \( \mathcal{G} = \{1, 2, \ldots, |\mathcal{G}|\} \), are directly connected to and managed by a central controller as shown in Fig. 1. The set, \( \mathcal{G} \), is a gNB sharing group in which multiple gNBs are largely overlapped in their radio coverage. For the highly overlapped gNBs in a sharing group, orthogonal resources are allocated among them, and RBs can
TABLE I: Summary of Important Notations

<table>
<thead>
<tr>
<th>Notation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$G$</td>
<td>set of gNBs in one sharing group</td>
</tr>
<tr>
<td>$\mathcal{Y}$</td>
<td>set of URLLC services</td>
</tr>
<tr>
<td>$\mathcal{Z}$</td>
<td>set of eMBB services</td>
</tr>
<tr>
<td>$G_g$</td>
<td>radio coverage area of gNB $g \in G$</td>
</tr>
<tr>
<td>$R_g$</td>
<td>radius of $G_g$</td>
</tr>
<tr>
<td>$(g, y)$</td>
<td>URLLC service $y \in \mathcal{Y}$ under gNB $g$</td>
</tr>
<tr>
<td>$(g, z)$</td>
<td>eMBB service $z \in \mathcal{Z}$ under gNB $g$</td>
</tr>
<tr>
<td>$A_{g,y}$</td>
<td>number of RBs allocated to service $(g, y)$</td>
</tr>
<tr>
<td>$\delta_y$</td>
<td>delay upper-bound for URLLC service $y$</td>
</tr>
<tr>
<td>$\mathfrak{N}_z$</td>
<td>average throughput lower-bound for eMBB service $z$</td>
</tr>
<tr>
<td>$v_z$</td>
<td>mobility speed of service $z$ nodes</td>
</tr>
<tr>
<td>$\tau_i$</td>
<td>pausing time at point $a_i$ for eMBB nodes</td>
</tr>
<tr>
<td>$D_{g,z}$</td>
<td>average data rate by one RB for eMBB service $(g, z)$</td>
</tr>
<tr>
<td>$D_{g,y}$</td>
<td>average data rate by one RB for URLLC service $(g, y)$</td>
</tr>
<tr>
<td>$A_l, \mu_l, \lambda_l$</td>
<td>high and low arriving rate for switched Poisson process (SPP)</td>
</tr>
<tr>
<td>$\epsilon_y$</td>
<td>maximal dropping probability for URLLC service $y$</td>
</tr>
<tr>
<td>$\psi_y$</td>
<td>maximal decoding error for URLLC service $y$</td>
</tr>
<tr>
<td>$\Phi_y$</td>
<td>reliability threshold for URLLC service $y$</td>
</tr>
<tr>
<td>$r_{g,z}$</td>
<td>node-to-gNB distance for service $(g, z)$</td>
</tr>
<tr>
<td>$E_{g,z}$</td>
<td>average number of required RBs for service $(g, z)$</td>
</tr>
<tr>
<td>$q_{g,y}$</td>
<td>URLLC traffic queued in the buffer for service $(g, y)$</td>
</tr>
<tr>
<td>$\delta_y$</td>
<td>queuing delay upper-bound for service $y$</td>
</tr>
<tr>
<td>$Q_{g,y}(x)$</td>
<td>the probability of queue $q_{g,y}$ exceeding $x$</td>
</tr>
</tbody>
</table>

be reused among different sharing groups. Time is divided into a sequence of fixed-duration time slots with length $t_0 = 1$ ms for eMBB service scheduling. Each time slot is further partitioned into a number of mini-slots with length $t_u \in [0.125, 1]$ ms for URLLC service scheduling.

For heterogeneous URLLC services and eMBB services, we leverage radio resource slicing to partition the physical network into multiple URLLC and eMBB slices. Each slice represents a customized service. Let set $\mathcal{Y}$ denote all the URLLC services with different QoS requirements, and set $\mathcal{Z}$ denote all eMBB services. Each gNB $g \in G$ needs to accommodate the service requests of nodes inside its coverage area $G_g$, which is simplified to a circle with radius $R_g$. Let $(g, y)$ indicate URLLC service $y \in \mathcal{Y}$ under gNB $g$, $(g, z)$ the eMBB service $z \in \mathcal{Z}$ under gNB $g$. The number of URLLC nodes served by service $(g, y)$ is denoted by $N_{g,y}$, and the number of eMBB nodes served by service $(g, z)$ is denoted by $N_{g,z}$. For network-level RB pre-allocation, the number of RBs allocated to service $(g, y)$ is denoted by $A_{g,y}$, with $g \in G, y \in \mathcal{Y}$. The unit of $A_{g,y}$ is RBs per mini-slot, i.e., $A_{g,y}$ RBs can be scheduled within one mini-slot. The gNBs are mutually connected via Xn interfaces [21]. At gNB-level RB scheduling, for any service $(g, y)$, a queue, $q_{g,y}$, with length $\delta_y A_{g,y}$, and a cache with size $C_{g,y}$ are maintained, where $\delta_y$ in mini-slots represents the delay upper bound of URLLC service $y$. The minimal average throughput requirement for eMBB service $z$ is denoted by $\mathfrak{N}_z$.

B. Mobility Model

Since URLLC nodes in IIoT scenarios are in general with low mobility [22], they are assumed to be quasi-static during a slicing period. The widely adopted RWP model is used to characterize each eMBB node’s movement within the coverage of a gNB [23]. For gNB $g$, an eMBB node initially locates at source point $a_0 \in G_g$, and then moves along a straight line from $a_0$ to a destination point $a_1 \in G_g$ with speed $v_1$. Both $a_0$ and $a_1$ are independently and uniformly distributed over $G_g$. The speed $v_z \in [v_{z,\text{min}}, v_{z,\text{max}}]$ for service $z$ follows certain probability density function (PDF). The trajectory of one eMBB node in $G_g$ is composed of multiple source-destination segments where the $i$th segment is denoted by $(a_{i-1}, a_i)$. We further define a pausing time, $\tau_i \in [0, T_z]$, for service $z$ at point $a_i$ to indicate the stop and wait time for nodes. The parameters $v_{z,\text{min}}, v_{z,\text{max}}$ and $T_z$ depend on the eMBB service types (e.g. autonomous vehicles, pedestrians, etc.) [23]. Note that the highly mobile eMBB nodes can make the traffic load of a gNB highly dynamic, which triggers frequent re-slicing of RAN resources and increases signaling overhead. We assume the eMBB node mobility is controlled such that the re-slicing overhead would not affect the overall network performance.

C. Communication Model

The average achievable throughput of eMBB slices can be obtained based on Shannon’s capacity [11]. While for URLLC services, the limited blocklength capacity theorem should be applied because the URLLC packet size is much smaller than that of conventional services [24]. For service $(g, z)$, the average data rate achieved by one RB is

$$D_{g,z} = W \log_2 (1 + \frac{P_i d_i m g_s^2}{N_0 W + \sum_{i \in I} I_i}). \quad (1)$$

For URLLC service $(g, y)$, we can calculate the maximal data rate achieved by one RB with decoding error $\psi_y$ [24]

$$D_{g,y} = W \log_2 (1 + \frac{P_i d_i m g_s^2}{N_0 W + \sum_{i \in I} I_i}) - \sqrt{\frac{V_{RB}}{W t_a f_Q^{\frac{1}{2}}(\psi_y)}}. \quad (2)$$

In (1) and (2), $W$ is the bandwidth of one RB, $P_i$ the transmit power of the gNB for one RB, $r_g$ the node-to-gNB distance, $N_0$ the noise power spectral density, $I_i$ the interference level from the interfering gNB $i$ using the same RB, $f_Q^{\frac{1}{2}}(\cdot)$ the inverse function of the $Q$-function, and $V_{RB}$ the channel dispersion of one RB, given by $V_{RB} = \frac{1}{(\ln 2)^2} \left[ 1 - (1 + P_i d_i m g_s^2)^{-2} \right]$ with $d_i$ and $m_i$ being the large-scale and small-scale path loss components of channel gain, respectively.

Without loss of generality, we assume that $W$, $P_i$, $d_i$ and $N_0$ are constants. For an RB, the interference from gNBs outside the sharing group is denoted by $\sum_{i \in I} I_i$, where $I$ is the set of interfering gNBs. To ensure the reliability requirements, the worst-case total interference among all RBs in one gNB is used to determine $D_{g,y}$. The channel gain, $m_i$, is modeled as a random variable. Considering the IIoT scenario where the networking environment is relatively stable and URLLC nodes are quasi-static, the dynamics of multi-path propagation is limited and no Doppler spread is considered. On the other hand, the randomness of $m_i$ can be averaged out when calculating the average throughput performance of eMBB nodes. For simplicity, in this work we assume that $m_i$ is constant for all eMBB and URLLC nodes.
D. Data Traffic Model

1) URLLC Data Traffic: The SPP is leveraged to model the burstiness and auto-correlation of URLLC packet arrivals [25], [26]. The traffic arrival rate of service \((g, y)\) switches between a high-rate Poisson process with parameter \(\lambda_{H,y}\) and a low-rate Poisson process with parameter \(\lambda_{L,y}\) (\(\lambda_{L,y} \leq \lambda_{H,y}\)). The intervals of the high and low traffic states follow an exponential distribution with parameter \(\mu_{H,y}\) and \(\mu_{L,y}\), respectively. Under the SPP traffic model for homogeneous URLLC nodes, the aggregate data traffic.

![Fig. 2: The MMPP model of service \((g, y)\)'s aggregate data traffic.](image)

URLLC packets in the next time slot. Note that all eMBB data transmissions can be suspended to the next slot as extreme case where burst URLLC traffic occupies all RBs in one slot. To test the performance of the proposed scheme in terms of guaranteeing QoS of eMBB services, we assume that the aggregated eMBB traffic within each gNB is a bulk traffic which consumes all the remaining RBs of URLLC services.

IV. NETWORK-LEVEL RESOURCE PRE-ALLOCATION

A. Average RB Requirements for eMBB Services

Given the RWP mobility model of eMBB nodes, the average RB requirement for service \((g, z)\) can be calculated based on the distribution of node-to-gNB distance \(r_{g,z}\).

Considering the RWP model with pausing (node remains static) state, the probability density function (PDF) of \(r_{g,z}\) is calculated as

\[
f(r_{g,z}) = p_{\tau} f_{\tau}(r_{g,z}) + (1 - p_{\tau}) f_{m}(r_{g,z})
\]

where \(p_{\tau}\) is the node pausing probability, \(f_{\tau}(r_{g,z})\) and \(f_{m}(r_{g,z})\) are PDFs of \(r_{g,z}\) when the node is in pausing and moving states, respectively. By representing pausing point \((r_{g,z}, \phi)\) in polar coordinates, we calculate \(f_{\tau}(r_{g,z})\) and \(f_{m}(r_{g,z})\) as

\[
f_{\tau}(r_{g,z}) = \int_{0}^{2\pi} \frac{1}{\pi R_g^2} r_{g,z} \text{d}\phi = \frac{2r_{g,z}}{R_g^2}, \quad r_{g,z} \in [0, R_g]
\]

\[
f_{m}(r_{g,z}) = \int_{0}^{2\pi} \frac{2 \phi^2 - \frac{r_{g,z}^2}{R_g^4}}{\pi R_g^4} \text{d}\phi, \quad r_{g,z} \in [0, R_g].
\]

The pausing probability, \(p_{\tau}\), is defined as the percentage of time that a node pauses during a long-running RWP process. Considering that a node pauses for \(\tau_i\) time period at destination \(i\), we have

\[
p_{\tau} = \lim_{T_i \to \infty} \frac{\sum_{i=1}^{T_i} \tau_i}{i \tau_i} = \frac{E(\tau_i)}{E(\tau_i) + E(\tau_{m,i})}
\]

where \(\tau_{m,i}\) represents the transition time for the node moving from point \(a_{i-1}\) to destination \(a_i\). The two random variables, \(\tau_i\) and \(\tau_{m,i}\), are independent. For the circular coverage area \(G_g\) with radius \(R_g\), we have

\[
E(\tau_i) = T_z/2
\]

\[
E(\tau_{m,i}) = \frac{128}{45 \pi R_g^2} \ln \left( \frac{v_{\max}}{v_{\min}} \right) \frac{v_{\max}}{v_{\min}}.
\]

By substituting (5)-(7) into (4), the PDF of \(r_{g,z}\) can be derived as

\[
f(r_{g,z}) = \frac{90nT_z r_{g,z} (2R_g^2 - R_{g,z}^2)}{(45nT_z + 256R_g^4V)R_g^4} + \frac{4r_{g,z}}{R_g^3} - \frac{R_{g,z}^3}{R_g^4}.
\]
where
\[ K = \frac{P_t d m_t}{N_0 W + \sum_{i \in I} I_i} \] (10a)
\[ A = \frac{\frac{45\pi T_c}{2} + 256 R_g V}{90\pi T_c} - \frac{4}{R_g^2} \] (10b)
\[ C = \frac{4}{R_g^2} - \frac{\frac{45\pi T_c}{2} + 256 R_g V}{90\pi T_c} R_g^2. \] (10c)

Therefore, the average number of required RBs for eMBB service \((g,z)\) is
\[ E_{g,z} = N_{g,z} \left[ \frac{R_z}{E(D_{g,z})} \right]. \] (11)

B. Delay Violation Probability for URLLC Services

Within a mini-slot, the aggregate URLLC traffic load for service \((g,y)\) follows the MMPP. The pre-allocated \(A_{g,y}\) RBs are used to transmit the packets. By normalizing the URLLC traffic in the unit of RBs per mini-slot, the downlink URLLC transmission for service \((g,y)\) can be modeled as a multi-state MMPP/D/1 queue, \(q_{g,y}\), where the deterministic service time equals \(1/A_{g,y}\) mini-slots. The length of \(q_{g,y}\) represents the URLLC traffic queued in the buffer. Since \(A_{g,y}\) RBs are available in each mini-slot, the maximal length for \(q_{g,y}\) should be limited to \(\delta_y A_{g,y}\) where \(\delta_y\) is the allowed maximal queuing delay of service \((g,y)\).

To formulate the network-level RB pre-allocation problem, the probability that queue length \(q_{g,y}\) exceeding \(x\), \(Q_{g,y}(x)\), should be calculated first. For computation complexity, we consider the following Laplace-Stieltjes transformation of the MMPP/G/1 queue’s delay:
\[ \mathcal{D}(s) = s(1 - \rho) \mathbf{g} \left[ \mathbf{I} + \mathbf{R} - \mathbf{L}\left(1 - \mathcal{L}(s)\right) \right]^{-1} \mathbf{e} \] (12)
where \(\mathcal{L}(s)\) is the Laplace-Stieltjes transformation of service time, \(\rho\) the ratio of the mean traffic rate of MMPP to the steady-state vector defined in [27], \(\mathbf{I}\) the identity matrix, \(\mathbf{R}\) and \(\mathbf{L}\) the transition rate matrix and the arrival rate matrix of MMPP, respectively. For deterministic service time \(1/A_{g,y}\), we have \(\mathcal{L}(s) = e^{-s/\overline{r}_y}\) by transforming \(Q_{g,y}(x)\) to delay survivor function \(Q_{g,y}(t A_{g,y})\) [29]. (12) is equivalent to the Laplace-Stieltjes transformation of queue length.

Based on mathematical studies of MMPP/G/1 queue [29] [30], \(Q_{g,y}(x)\) equals the summation of multiple exponential terms. The term with the largest negative exponential factor dominates the slop of \(Q_{g,y}(x)\). Accordingly, \(Q_{g,y}(x)\) can be approximated by the single exponential function [29]
\[ Q_{g,y}(x) = p_0 e^{s_y x} \] (13)
where \(s_y\) is the largest negative root of the denominator of (12), and \(p_0\) is the probability of a non-empty queue. By approximating \(\mathcal{L}(s)\) with its first three Maclaurin series components, i.e., \(e^{-s/\overline{r}_y} = 1 - \frac{s}{\overline{r}_y} + \frac{b_2 s^2}{2}\), \(s_y\) can be calculated from the largest negative roots of
\[ \det \left[ \frac{d}{d} s I + \mathbf{R} - \mathbf{L}(s - \frac{s_y^2}{2}) \right] = 0. \] (14)

To determine \(p_0\), we implement the asymptotic approximation method to scale down the \((N + 1)\)-state MMPP into a 2-state MMPP [30]. For an \((N + 1)\)-state MMPP/D/1 queue, the states \{\(M+1, M+2, \ldots, N\) and \{0, 1, \ldots, M\} are treated as overlord (OL) states and underload (UL) states, respectively, where \(M = \left[ \frac{N p_{on}}{\rho} \right]\). With the 2-state MMPP, we have
\[ p_0 = \left(\frac{\lambda_{OL}}{\lambda_{OL} + \mu_{UL}}\right) \sum_{k=0}^{N} \pi_{k} \] (15)
where \(\lambda_{OL}, \mu_{UL}\), and \(\lambda_{UL}\) are parameters of the 2-state MMPP as given in [30].

C. Problem Formulation

The objective of the network-level RB slicing problem is to pre-allocate the minimal number of RBs, denoted by \(A_{g,y}\), for all URLLC services at all gNBs in one sharing group, while satisfying their delay and reliability requirements. For transmission scheduling of eMBB services after that of URLLC services, the average number of required RBs for service \((g,z)\) is constrained by
\[ \sum_{z \in Z} (1 - \rho_{g,y}) A_{g,y} \geq \sum_{z \in Z} E_{g,z}, \] (16)
where \(\rho_{g,y}\) is the mean URLLC traffic rate normalized by the number of pre-allocated RBs, \(A_{g,y}\).

Given the delay upper bound \(\delta_y\) and the pre-allocated RBs, \(A_{g,y}\), at each mini-slot, the delay requirement of URLLC service \((g,y)\) is ensured by limiting the maximal queue length \(\delta_y A_{g,y}\). The transmission of overflowed URLLC packets is scheduled by temporarily accessing available RBs from other services different from service \((g,y)\), i.e., \((g',y') \in (\mathcal{Y} \times \mathcal{G} \setminus \{(g,y)\})\). Note that any other service can be a different service \(y' \in \mathcal{Y}\) on the same gNB, \(g' \in \mathcal{G}\), and different service \(y' \in \mathcal{Y}\) on different gNBs. To meet the URLLC reliability requirement, the probability that the other services have insufficient available RBs to support the exceeded traffic load should be smaller than \(\epsilon_y\). Given \(Q_{g,y}(x)\), let \(E_{g,y}\) represent the number of RBs that service \((g,y)\) needs to borrow from other services in a mini-slot. Let \(P_{g',y'}(A_{g',y'}, X_{g,y})\) denote the probability that any other service has larger than or equal to \(E_{g,y}\) available RBs to be accessed by service \((g,y)\) at one mini-slot. The probability can be expressed as
\[ P_{g',y'}(A_{g',y'}, X_{g,y}) = \left(1 - Q_{g',y'}(0)\right) \times \left(\sum_{k=0}^{\overline{N}_{g',y'}} \mathcal{P}_{k,g',y'}(A_{g',y'} - X_{g,y}) \pi_k\right) \] (17)
where \((1 - Q_{g',y'}(0))\) is the probability that the queue length service \(y'\) \(q_{g',y'}\) is zero at the beginning of a mini-slot, \(\sum_{k=0}^{\overline{N}_{g',y'}} \mathcal{P}_{k,g',y'}(A_{g',y'} - X_{g,y}) \pi_k\) the probability that the newly arrived traffic load of service \((g',y')\) in a mini-slot is smaller than \((A_{g',y'} - X_{g,y})\). Note that \(\mathcal{P}_{k,g',y'}(\cdot)\) is the cumulative distribution function (CDF) of Poisson distribution at the \(k\)th state of MMPP for service \((g',y')\). The distribution of \(X_{g,y}\) varies with \(A_{g',y'}\). To study its impact, we consider two cases, i.e., \(0 < X_{g,y} \leq C_{g,y}\) and \(X_{g,y} > C_{g,y}\). Note that \(C_{g,y}\) represent the size of a cache space to store overflowed data from service \((g,y)\), which is normalized to the number of
required RBs using (3). When \( X_{g,y} > C_{g,y} \), the packet loss is unavoidable due to exceeding maximal cache size. The unavoidable packet dropping probability \( \gamma_{g,y} = Q_{g,y}(C_{g,y} + \delta_y A_{g,y}) \), i.e., the probability that the length of \( q_{g,y} \) exceeds \( (C_{g,y} + \delta_y A_{g,y}) \). For \( 0 < X_{g,y} \leq C_{g,y} \), the dropping probability due to insufficient available RBs from other gNBs is given by

\[
E(A_{g,y}, A_{g',y'}, C_{g,y}) = \prod_{g',y'} [1 - P_{g',y'}(A_{g',y'}, C_{g,y})] \times \left[ Q_{g,y}(\delta_y A_{g,y}) - \gamma_{g,y} \right] \tag{18}
\]

where \( \prod_{g',y'} [1 - P_{g',y'}(A_{g',y'}, C_{g,y})] \) denotes the probability that the number of available RBs of all other services is less than \( C_{g,y} \). By ensuring that the dropping probability is not larger than \( (e_y - \gamma_{g,y}) \), we formulate the optimization problem, with the objective of minimizing the total number of pre-allocated RBs for all services on all gNBs in a sharing group:

\[
\min_{A_{g,y}, g \in G, y \in Y} \sum_{g \in G} \sum_{y \in Y} A_{g,y} \tag{19}
\]

\[s.t. \quad E(A_{g,y}, A_{g',y'}, C_{g,y}) \leq e_y \tag{19a}\]

\[A_{g',y'} \geq C_{g',y'} \tag{19b}\]

\[\gamma_{g,y} \leq e_y \quad \forall g, y \tag{19c}\]

\[\sum_{y \in Y} (1 - \rho_{g,y}) A_{g,y} \geq \sum_{z \in Z} E_{g,z}, \quad \forall g \tag{19d}\]

In (19), constraint (19a) is for the reliability of any URLLC services at any gNB. The minimal number of pre-allocated RBs for service \((g',y')\) is constrained by (19b), which indicates that any \( A_{g',y'} \) should always be not smaller than \( C_{g',y'} \) to ensure sufficient available RBs. The variable ranges for (19a) and (19b) are the same, i.e., \( \forall g, y, (g', y') \in \{G \times Y\} \setminus (g, y) \). Constraint (19c) implies that the minimal number of RBs pre-allocated to service \((g, y)\) must ensure that \( \gamma_{g,y} \) is not larger than \( e_y \), given \( C_{g,y} \). The minimal long-term average throughput for eMBB services is satisfied by (19d).

Since the optimization variable, \( A_{g,y} \), is an integer variable, and (19a) involves numerical calculations of matrix determinant’s roots and eigenvalues, problem (19) is an INLP. From (19a), the dropping probabilities for service \((g, y)\) decrease monotonically as the \( A_{g,y} \) increases, when other \( A_{g',y'} \) keep constant. Therefore, problem (19) can be categorized as a Knapsack problem which optimizes the combination of \( A_{g,y} \) for all URLLC services to minimize the total cost (summation of \( A_{g,y} \)), while ensuring constraints (19a)-(19d) [19]. Therefore, (19) is at least NP-complete.

D. Network-level RB Pre-allocation Algorithm

To solve the NP-complete problem (19) in polynomial time, a heuristic algorithm with reduced complexity is proposed. The algorithm initially assigns over-provisioned RBs to each service \((g, y)\) which satisfy reliability requirements for all URLLC services. Then, the algorithm iteratively reduces the amount of RBs assigned to each service until violating constraints (19a)-(19d). By reduce the same amount of RBs at each iteration, different service can cause different amounts of dropping probability increment for the RAN. Therefore, the proposed algorithm is designed to reduce maximal number of RBs before violating the upper bound of dropping probability (i.e., URLLC reliability requirement \( \delta_y \)), which approximates the minimal amount of RBs to satisfy URLLC QoS requirements in the RAN. Details of the algorithm are presented in Algorithm 1.

Initialization: The algorithm starts with assigning \( A_{g,y,max} \) RBs to service \((g, y)\) by letting \( Q_{g,y}(\delta_y A_{g,y,max}) = e_y \), and initializes \( \eta_{g,y} \) which equals the dropping probability of service \((g, y)\) with \( A_{g,y,max} \) (Line 1–2).

Dropping probability updates: Given the initialized variables, the RB pre-allocation algorithm proceeds its main loop iteratively (Line 3–23). For service \((g, y)\), we reduce its pre-allocated RB numbers at the \( j \)th iteration, \( A_{g,y}[j] \), by one RB to obtain \( A_{g,y}^{temp}[j + 1] \), and update the temporary dropping probabilities of all URLLC services, i.e., \( \eta^{temp}_{g,y} \) for service \((g, y)\) and \( \eta^{temp}_{g',y'} \) for other services (Line 4–8).

Optimal RB pre-allocation update per iteration: Define \( \xi_{g,y} \) as the amount of dropping probability increase when \( A_{g,y}[j] \) is reduced by one RB in a iteration, a smaller \( \xi_{g,y} \) indicates a less increase of dropping probability which allows more potentials for further RB amount reductions. Calculating \( \xi_{g,y} \) for all services \((g, y), \forall g \in G, y \in Y \) (Line 9–17). By choosing service \( y^* \) under gNB \( g^* \) which achieves the minimal \( \xi_{g,y} \), the algorithm updates \( A_{g,y}[j + 1] \) as \( A_{g,y}^{temp}[j + 1] \), and all services’ dropping probabilities (Line 18–22).

Iteration stop condition: The algorithm keeps searching the possible solution space \( \Omega \) and stops when no RBs in \( A_{g,y}[j] \) can be further reduced without violating constraints (19a)-(19d) (Line 11–12, 15–17).

Complexity analysis: We compare the computational complexity between our proposed algorithm and a brute-force (BF) search algorithm which can find the optimal solution for (19). Assume that the size of \( \Omega \), the computation complexity of BF search can be estimated as \( O(|G| \times |Y|) \), which increases exponentially with \(|G| \times |Y|\) times. Given the total number of iterations \( K \), the computation complexity of the proposed heuristic is estimated as \( O(K(|G| \times |Y|^2)) \), which increases quadratically (or linearly) as \(|Y| \) or \(|G| \) increases. Since \( K \) is in the same order as \(|\Omega| \), the computation complexity of the proposed heuristic algorithm is much lower than that of the BF search. Simulation results also indicate that the performance gap between the two algorithms are close (less than 5%), especially for the 2-gNB and 3-gNB cases where the gap almost vanishes.

V. gNB-level RB Scheduling

Conventional RB scheduling algorithms, e.g., the enhanced proportional fair (EPF), are not suitable for ensuring the strict reliability and latency requirements of URLLC services [20]. Meanwhile, the inter-service and inter-gNB RB sharing, which are essential to increasing multiplexing gain, are not considered in those algorithms. Given \( A_{g,y} \) number of pre-allocated RBs, we design a gNB-level RB scheduling scheme for URLLC service \((g, y)\) to ensure QoS and support inter-gNB RB sharing. For the eMBB traffic which is scheduled
using the remaining RBs after the scheduling of all URLLC traffic, the EFP algorithm is applied.

Let \( q_{g,y}(t) \) denote the queue length (normalized to number of RBs) at mini-slot \( t \). Upon URLLC packet arrivals \( X_{g,y}(t) \) for service \((g, y)\) at mini-slot \( t \) (normalized to number of RBs), the RB scheduling scheme includes the following three states depending on the values of \( X_{g,y}(t) \), \( A_{g,y} \), and \( q_{g,y}(t) \) at the beginning of mini-slot \( t \):

1) In **State 1**: When \( X_{g,y}(t) + q_{g,y}(t) \leq A_{g,y} \), the pre-allocated RBs are sufficient to support the transmission of both newly arrived and queued URLLC packets at \( t \). All \( (X_{g,y}(t) + q_{g,y}(t)) \) RBs are directly scheduled to support the packet transmission in mini-slot \( t \) + 1.

2) In **State 2**: When \( A_{g,y} < X_{g,y}(t) + q_{g,y}(t) \leq \delta_y A_{g,y} \), the maximal RBs that can be scheduled to service \((g, y)\) in mini-slot \( t \) are fully occupied, but the queue length threshold is not exceeded. In this state, the gNB schedules \( A_{g,y} \) RBs for data packet transmission in mini-slot \( t \) + 1 according to first-in-first-out (FIFO), and the remaining data with the length of \( q_{g,y}(t + 1) = X_{g,y}(t) + q_{g,y}(t) - A_{g,y} \) stay in the queue.

3) In **State 3**: When \( X_{g,y}(t) + q_{g,y}(t) > \delta_y A_{g,y} \), the scheduling queue capacity is exceeded. The amount of overflowed data packets, denoted by \( (X_{g,y}(t) + q_{g,y}(t) - \delta_y A_{g,y}) \) is cached, and a broadcast message requesting RBs from other services in the sharing group is sent. By caching the overflowed data, the remaining data in the queue are scheduled and transmitted following the same procedure in Case 2. At mini-slot \( t + \delta_y \), service \((g, y)\) receives the information regarding the amount of available RBs of size \( w \) from other services. Then, at mini-slot \( t + \delta_y + 1 \), gNB \( g \) uses local RBs to schedule the transmission of \( A_{g,y} \) amount of data, and temporarily borrows \( w \) RBs to schedule the transmission of the remaining data with the size \( (X_{g,y}(t) + q_{g,y}(t) - \delta_y A_{g,y}) \). The detail resource sharing rules can vary in real implementations. We consider a rule that all \( w \) RBs should come from the service with the maximum number of available RBs.

Through the inter-gNB resource sharing (supported by the Xn wired links between gNBs [21], the proposed gNB-level RB scheduling scheme can achieve the collision-free RB scheduling. Since the amount of messages exchanged between gNBs are limited and transmitted via wired links, the packet loss due to inter-gNB communication can be neglected.

VI. SIMULATION RESULTS

We conduct simulations to evaluate the performance of the proposed two-level RAN soft-slicing scheme. For the reliability requirements of URLLC services, we assign the same value to \( e_y \) and \( \psi_y \) with their summation equivalent to service \( y \)’s packet loss upper bound \( \Phi_y \) [25], i.e., \( \Phi_y = 1 \times 10^{-5} \) and \( e_y = \psi_y = \Phi_y/2 = 5 \times 10^{-6} \). Without loss of generality, for each gNB, we set 33 dBm transmit power, \( R_g = 100 \) m, \( N_0 = -174 \) dBm/Hz, \( C_g = 60 \) RBs, 3 interfering gNBs, two types of eMBB nodes (high and low mobility), and the mean value of large-scale and small-scale path loss components \( d_l \) and \( m_l \) as 1.0. The node-to-gNB distances and node-to-interfering gNB distances of all nodes in one gNB are uniformly distributed between [10, 100] m and [250, 500] m, respectively. Given the 0.125 ms URLLC scheduling interval, the bandwidth \( W \) for one RB is 180 kHz [31]. The default delay upper bound \( \delta_y \) for all URLLC services is set to be 3 mini-slots (i.e., 0.375 ms), which is more strict than the 1 ms URLLC latency requirement. The minimal requirement of eMBB average throughput is 40 RBs per mini-slot.

We first evaluate the performance of the proposed soft slicing scheme in terms of QoS guarantee for URLLC (packet loss ratio) and eMBB (average throughput) services in a 3-gNB scenario. Detailed parameters of each service and gNB are shown in Table II. For each gNB, traffic are generated for the duration of 5 \times 10^5 mini-slots. The aggregated average throughput requirement for all eMBB services is 240 Mbps per gNB. It can be seen from Table II that the packet drop ratios of all services in different gNBs are lower than their packet loss upper bounds, which indicates that our two-level soft RAN slicing scheme is able to ensure the reliability constraint for URLLC services. The aggregated eMBB average throughput for every gNB is higher than the threshold 240 Mbps. Fig. 3 shows the cumulative distribution functions (CDFs) of different URLLC services’ traffic scheduling queue lengths. It is noted that the ratio of packet losses in different scenarios are all guaranteed smaller than the upper bound \( \Phi_y \).
TABLE II: Simulation parameters and resulting performance metrics

<table>
<thead>
<tr>
<th>Scenario Parameters</th>
<th>gNB 1</th>
<th>gNB 2</th>
<th>gNB 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Services</td>
<td>Service 1</td>
<td>Service 2</td>
<td>Service 3</td>
</tr>
<tr>
<td>$\lambda_e$ (Packets/mini-slot)</td>
<td>50, 0</td>
<td>50, 8</td>
<td>65, 0</td>
</tr>
<tr>
<td>$1/\mu_e$, $1/\mu_e$ (mini-slots)</td>
<td>1, 15</td>
<td>1, 5</td>
<td>1, 10</td>
</tr>
<tr>
<td>URLLC node number</td>
<td>4</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>Delay upper bound $\delta_y$ (mini-slots) [32]</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Packet loss upper bound $\Phi_y$ ($\times 10^{-9}$) [32]</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Performance of the proposed two-level soft-slicing scheme

| Pre-allocated RBs number | 118 | 114 | 158 | 86 | 71 |
| Scheduled packets number | 7219093 | 27081325 | 20967354 | 7234467 | 7217343 |
| Dropped packets number   | 10 | 86 | 27 | 4 | 118 |
| Packets drop ratio ($\times 10^{-5}$) | 0.139 | 0.318 | 0.129 | 0 | 1.635 |
| RB request probability ($\times 10^{-4}$) | 0.24 | 0.88 | 0.34 | 0.48 | 4.36 |
| RB sharing probability ($\times 10^{-3}$) | 3.88 | 0.14 | 0.16 | 0.96 | 0.16 |
| eMBB average throughput (Mbps) | 756.4635 | 888.2135 | 270.09 |

Figure 3: The CDF of RB amount in different URLLC service queues.

Fig. 4: Average number of RBs per gNB as $\epsilon_y$ and $\delta_y$ varies.

allowed [20], [25]. We simulate 3-gNB and 5-gNB scenarios where each gNB runs one URLLC service with the same parameter set the same as Service 1 in Table II. Fig. 4 shows the RBs requirements of the hard-slicing scheme and the proposed soft-slicing scheme under different reliability and delay constraints. In Fig. 4, it can be seen that the hard-slicing scheme consistently requires more RBs than the proposed soft-slicing scheme. The gaps become larger as $\epsilon_y$ becomes smaller, which demonstrates the advantage of the proposed scheme in terms of ensuring the strict reliability constraints for URLLC services.

Figure 5 shows the impact of URLLC node number and traffic arriving rate $\lambda_L$ upon the performance of soft and hard slicing. Simulation parameters for Service-1 and Service-3 are defined in II. Fig. 5(a) shows that the average number of RBs per gNB increases with the number of URLLC nodes, while the proposed scheme can always reduce RB consumption as the hard-slicing for all services. In Fig. 5(b), the average number of RBs per gNB increases with the traffic arriving rate indicator $\lambda_L$. Moreover, the performance gap between the soft and hard slicing decreases as $\lambda_L$ increases, because a high traffic load leads to a reduced number of available RBs for sharing. Specifically, when $\lambda_L > 8$ packets/mini-slot, the available RBs for eMBB services decreases significantly. Therefore, more RBs are required to support the minimal average throughput of the eMBB service, which causes an upward trend of Service-3’s curve in Fig. 5(b).

VII. CONCLUSION

In this paper, a two-level soft RAN slicing scheme has been proposed to enable dynamic radio resource sharing among different network slices. To guarantee differentiated QoS requirements of URLLC and eMBB services in IIoT scenarios, we formulate a network-level RB pre-allocation problem by considering both the QoS requirements and the inter-gNB resource sharing probabilities in optimizing the number of pre-allocated RBs among gNBs. Due to the NP completeness of the formulated problem, a low-complexity heuristic algorithm has been proposed. Then, given the pre-allocated resources, a collision-free gNB-level RB scheduling scheme has been designed to enable URLLC devices to temporarily access other
available gNB resources in a mini-slot level for QoS guarantee. Simulation results have demonstrated the effectiveness of the proposed scheme in terms of differentiated QoS provisioning for the coexistence of URLLC and eMBB services, and the improved multiplexing gain compared with the “hard-slicing” scheme. This study can be extended by incorporating machine learning methods. A learning-based soft RAN slicing solution should be investigated to address uncertain traffic that cannot be properly characterized by mathematical models.

REFERENCES


[31] 3GPP TS 38.211 V15.0.0, “NR: Physical channels and modulation,” 2017.