sensors

MDPI

Article

Joint User-Slice Pairing and Association Framework Based on H-NOMA in RAN Slicing

Mai A. Riad [1,*], Osama El-Ghandour [1,2] and Ahmed Abd El-Haleem [1]

[1] Electronics and Communications Engineering Department, Faculty of Engineering, Helwan University, Helwan 11792, Egypt
[2] Electronics and Communications Engineering Department, Faculty of Engineering, New Cairo Academy University, Cairo 11865, Egypt
* Correspondence: mai_awad_reyad@h-eng.helwan.edu.eg

Abstract: Multiservice cellular in Radio Access Network (RAN) Slicing has recently attained huge interest in enhancing isolation and flexibility. However, RAN slicing in heterogeneous networks (HetNet) architecture is not adequately explored. This study proposes a pairing-network slicing (NS) approach for Multiservice RAN that cares about quality of service (QoS), baseband resources, capacities of wireless fronthaul and backhaul links, and isolation. This intriguing approach helps address the increased need for mobile network traffic produced by a range of devices with various QoS requirements, including improved dependability, ultra-reliability low-latency communications (uRLLC), and enhanced broadband Mobile Services (eMBB). Our study displays a unique RAN slicing framework for user equipment (UE) for joint user-association. Multicell non-orthogonal multiple access (NOMA)-based resource allocation across 5G HetNet under successive interference cancelation (SIC) is seen to achieve the best performance. Joint user-slice pairing and association are optimization problems to maximize eMBB UE data rates while fulfilling uRLLC latency and reliability criteria. This is accomplished by guaranteeing the inter- and intra-isolation property of slicing to eliminate interferences between eMBB and uRLLC slices. We presented the UE-slice association (U-S. A) algorithm as a one-to-many matching game to create a stable connection between UE and one of the base stations (BSs). Next, we use the UE-slice pairing (U-S. P) algorithm to find stable uRLLC-eMBB pairs that coexist on the same spectrum. Numerical findings and performance analyses of the submitted association and pairing technique show they can all be RAN slicing criteria. We prove that the proposed algorithm optimizes system throughput while decreasing uRLLC latency by associating and pairing every uRLLC user in mini slots.

Keywords: 5G; RAN slicing; HetNet; UE-slice association; UE-slice pairing; eMBB; H-NOMA; uRLLC; matching game

Citation: Riad, M.A.; El-Ghandour, O.; Abd El-Haleem, A. Joint User-Slice Pairing and Association Framework Based on H-NOMA in RAN Slicing. *Sensors* **2022**, *22*, 7343. https://doi.org/10.3390/s22197343

Academic Editors: Peter Han Joo Chong and Omprakash Kaiwartya

Received: 1 August 2022
Accepted: 22 September 2022
Published: 27 September 2022

1. Introduction

The tremendous recent advances in mobile communication, intelligent user devices, and traffic demand have necessitated cell density optimization in extensive networks to provide high data throughput and low latency [1]. Ultra-dense networks have been considered a promising solution for the fifth generation (5G) networks as network densification can boost network coverage and capacity while reducing operational and capital expenditures. In addition, the backhaul network needs to be enhanced to transmit traffic from base stations (BSs) to the core network (CN) and vice versa, and sub-6 GHz wireless backhaul (WBH) facilitates non-line-of-sight (NLOS) transmission, while minimal delay on radio access and backhaul lines has become crucial for delivering various services and applications, such as Voice over Internet Protocol (VoIP) and online gaming, with an acceptable quality of service in future cellular networks, e.g., VoIP and online gaming (QoS) [2]. However, traditional HetNets dedicate radio resources to UE in time or frequency domains,

i.e., orthogonal multiple access (OMA), where the availability of the radio resources strictly limits the number of served UEs at a given time instant. Considering the expected explosive number of devices, massive connectivity necessitates more spectrum-efficient access schemes with extended coverage [3].

On the other hand, non-orthogonal multiple access (NOMA) has recently attracted attention by permitting sharing of the same radio resources among a set of UEs. NOMA can support more users than the number of available orthogonal resources [4], leading to higher spectral efficiency and user fairness compared to standard OMA techniques. The principle of NOMA leverages the concept of superposition coding (SPC) at the transmitter to multiplex users in the power domain and successive interference cancellation (SIC) at the receiver [5].

The 5G network services have been classified into three categories by the International Telecommunication Union (ITU): Enhanced Mobile Broadband (eMBB), Ultra-reliable and Low-latency Communications (uRLLC), and Massive Machine Type Communications (mMTC). The eMBB service focuses on high bandwidth requirement services, uRLLC focuses on latency-sensitive services, and mMTC focuses on services with high connection density requirements [6]. However, those services cannot always be achieved through a common network setting. To allow the coexistence of these heterogeneous services with diverse requirements within the same Radio Access Network (RAN) architecture, the concept of network slicing (NS) has been proposed [7], which slices the network into logical and physical sub-networks, usually with customized requirements in terms of latency, energy efficiency, mobility, massive connectivity, and throughput, aiming at guaranteeing minimum performance requirements and isolation [8]. The network slice comprises RAN and core network [9] to support end-to-end service requirements. This can be performed thanks to network softwarization and virtualization (SDN, NFV), which are considered the primary enabler of Resource as a Service (RaaS) beyond 5G (B5G) [10]. As for RAN slicing, additional challenges are faced due to limited capacity, radio channels, and performance isolation, which is still under development [11].

The uRLLC service is intended for event-driven, mission-critical, and industrial settings that may be helped in achieving QoS criteria such as ultra-low latency and ultra-high dependability. In addition, the standard uRLLC sets rigorous latency and reliability standards, with typical latency and reliability requirements of 1 ms/packet and up to 99% successful packet delivery [12]. Since both the eMBB and the URLLC are critical components of communication traffic in 5G, several researchers have addressed the problem of the coexistence between various services [13]. Some of the main challenges in HetNets are the user association process and the user pairing process. The users must be associated to different base stations (BSs) in the HetNets [14].

Consequently, and motivated by the aforementioned benefits of NOMA and RAN slicing, this paper considers H-NOMA in RAN slicing and studies the problem of the coexistence of services with heterogeneous requirements. User association and pairing algorithms are considered the various types of UEs and their different connection demands in a HetNet.

1.1. Related Work

Wireless networks emphasize resource slicing to increase spectrum utilization from NS problems in HetNet, by including user association and resource allocation. Furthermore, most literature saw NS's constraint-isolated resource allocation as an issue. In [15] Earliest Deadline First (EDF) scheduling, originally used in real-time operating systems, is exploited for radio resource allocation in RAN slicing for the first time.

Multiplexing the traffic from URLLC and eMBB users constitutes a significant challenge in 5G, hence it has been tackled several times in the literature very recently. Diverse QoS demands, in addition to the spectrum scarcity problem call for innovative approaches in addressing the resource management problem. In [16], a communication-theoretic model involving non-orthogonal sharing of RAN resources in uplink network slicing is studied

with the three types of heterogeneous services such as eMBB, uRLLC, and mMTC, and the authors refer the approach to as H-NOMA. For details, see [17]; NOMA is an excellent choice for 5G services that have great spectrum efficiency (eMBB), comprehensive device connectivity (mMTC), and low transmission latency (uRLLC). The objective [18] is to simplify SIC while improving its value and utility for UE. In [19,20], a comprehensive survey of different candidate NOMA schemes for 5G is presented. In [21], rate-splitting multiple access (RSMA) is used for uRLLC transmission, where a uRLLC device separates its message into two sub-messages depending on the average signal-to-noise ratio (SNR) without immediate channel status information (CSI). The researchers investigated uRLLC and eMBB coexistence under a puncturing approach in multiple-input multiple-output (MIMO) NOMA systems by dividing the original issue into two sub-problems: user selection and power allocation in [22]. This article investigated the coexistence of eMBB and uRLLC services inside a cellular network with a reconfigurable intelligent surface (RIS) [23]. In [24], an enhanced Pre-emptive Scheduling (EPS) scheduler for joint URLLC, and eMBB traffic is proposed to extract the maximum possible eMBB ergodic capacity. In [25], eMBB and uRLLC communications were encoded in the cloud and cracked at the edge nodes to fulfil latency constraints in a Cloud radio access network (C-RAN) scenario. According to [26], eMBB transmission risk and uRLLC dependability are proposed as risk metrics for eMBB transmission to ensure a fair, proportionate allocation of resources to incoming uRLLC traffic. uRLLC scheduling is based on Deep Reinforcement Learning (DRL)-based learning to allocate traffic and optimize eMBB data rate and dependability [27]. This research uses linear, convex, and threshold eMBB rate loss to increase network resource usage and allocate the uRLLC traffic [28]. In [29], the authors overcome the co-channel problem of eMBB and uRLLC users by puncturing to improve the eMBB UE rate using a one-sided matching game. In [30], using a homogenous NOMA per slice in uplink and downlink with fairness enhances system sum-throughput for all users. The multi-connectivity (MC)-based strategy increases resources and minimizes wait times in [31] by spectrum sharing across various base stations and differential QoS for data and machine-to-machine (M2M) services in a dynamic network request.

There are many research efforts for better user association schemes suitable in NOMA for HetNets. In [32] power domain NOMA (PD-NOMA), game theory algorithms are proposed based on the QoS threshold to improve user association and power allocation. This work is presented by considering various case studies to demonstrate the effectiveness of PD-NOMA in ultra-dense networks. In [33], the unified PD-NOMA is utilized for user association in a dense heterogeneous network and improved user association, the overall system capacity, and throughput. In [34], the flexible user association is proposed in NOMA-based multiple base stations (BSs) networks to maximize the weighted sum rate of the system, and a user association optimization problem is formulated. While a NOMA user pairing scheme combined with the almost blank subframe (ABS) technology, and a dynamic NOMA power allocation scheme are presented to maximize the network fairness based on the optimized throughput of the edge users in [35]. But few writers focus on various services UEs' associations with BSs. This study is aimed to improve isolation and flexibility by resolving the UE association in 5G [36]. In [37], using a matching game, UE association in a cellular virtualization network captures UE and BS preferences, QoS needs, and backhaul limitations. The Pointer Network (PtrNet) architecture implements NOMA's joint UE pairing and association schemes [38]. Recently, a matching game has been proposed for wireless UE affiliation and resource allocation [39]. In [40], the authors intend to improve the eMBB network rate and decrease eMBB loss by convincing uRLLC UEs to cohabit with eMBB UEs by superposition readiness by matching game. In [41], matching game techniques for UE association in HetNets are presented, taking both downlink and uplink properties into account. Moreover, in [42], the optimization of downlink data rates was achieved with Radio access technology (RAT) heterogeneity by ensuring fairness and lowering uplink power usage. However, none of those mentioned works above considers

intra and inter-isolation problems between slices when analyzing cell association and coexistence between uRLLC and eMBB slices.

1.2. Contribution

Utilizing the recent prosperity of using the matching theory to solve combinatorial optimization problem is proposed for eMBB UEs/uRLLC UEs association and pairing in HetNets in this paper. A scenario where eMBB and URLLC UEs with different requirements can be associated with different BSs and pair with each other is investigated. The main point of this paper can be summed up as follows in Figure 1:

- Joint user-association and pairing algorithms are considered for eMBB and uRLLC users in HetNets while considering the different association and pairing metrics required for each user type.
- In our system architecture, eMBB users are treated as downlink (DL) data-rate-hungry users and uRLLC users as latency and reliability-constrained users. The user-slice association and pairing processes are expressed as a multi-objective optimization problem to optimize the DL data rate for UEs and reduce the DL transmitted latency for uRLLC UEs whereas considering the users' data rate-dependent QoS requirements and ensuring the intra-isolation for each slice by applying orthogonal frequency multiple access (OFDMA) technology between the similar service's users; meanwhile, the inter-isolation between slices is assigned a DL threshold rate for each slice.
- We suggest a system model in which eMBB traffic is transmitted over long TTIs $\mathcal{T}$, whereas uRLLC traffic is transmitted over short TTIs δ by superimposing the ongoing eMBB transmissions. Here, transmitting the incoming uRLLC traffic in the short TTI guarantees its delay demand. The data rate of eMBB traffic is picked up by Shannon's capacity considering the effect of uRLLC transmissions, while uRLLC depends on the finite block-length capacity model due to its small packet size nature.
- We separate the multi-objective optimization problem into UE-slice association and UE-slice pairing sub-problems. Moreover, we improve a framework based on a one-to-many matching game to find a solution for the UE-slice association sub-problem in which BSs and UEs rate each other based on clear preference measures that define UEs' and BSs' needs. Meanwhile, the UE-slice pairing sub-problem is optimized by a one-to-one matching game in which the associated uRLLC and eMBB UEs with the same BS rank one another through some mini slots. To our knowledge, no matching game has been researched in HetNet to solve the eMBB UEs and uRLLC UEs association and pairing problem.
- We represent utility functions for UEs and BSs that consider the UEs' different demands in terms of attained data rate, DL latency reliability, the threshold rate of each slice, dynamic criteria for limiting the number of attached UEs in HetNet, and the number of paired uRLLC UE with eMBB UEs (i.e., quota matching game), to achieve a reference rate for UEs.
- The joint UE-slice association and UE-slice pairing algorithms are proposed to solve this game. Then, the proposed algorithms are proven to arrive at a stable matching. As well as the computation complexity study for the suggested algorithms is given. Simulation results illustrate that our submitted algorithms outperform the comparable approaches, especially regarding DL latency performance for uRLLC UEs and DL rate performance for eMBB and uRLLC UEs. To illustrate the effectiveness of the proposed algorithms, we also compare their performance with other search schemes.

1.3. Organization

According to the following structure, the rest of this document is: Section 2 explains the models of the system. Section 3 explains the problem formulation, while Section 4 is performance analysis which shows a UE-slice association solution and UE-slice pairing solution. Section 5 is dedicated to numerical simulation and performance analysis. Further-

more, future work is discussed in Section 6. Finally, concluding remarks are presented in Section 7 The abbreviations used in this paper are given at the end.

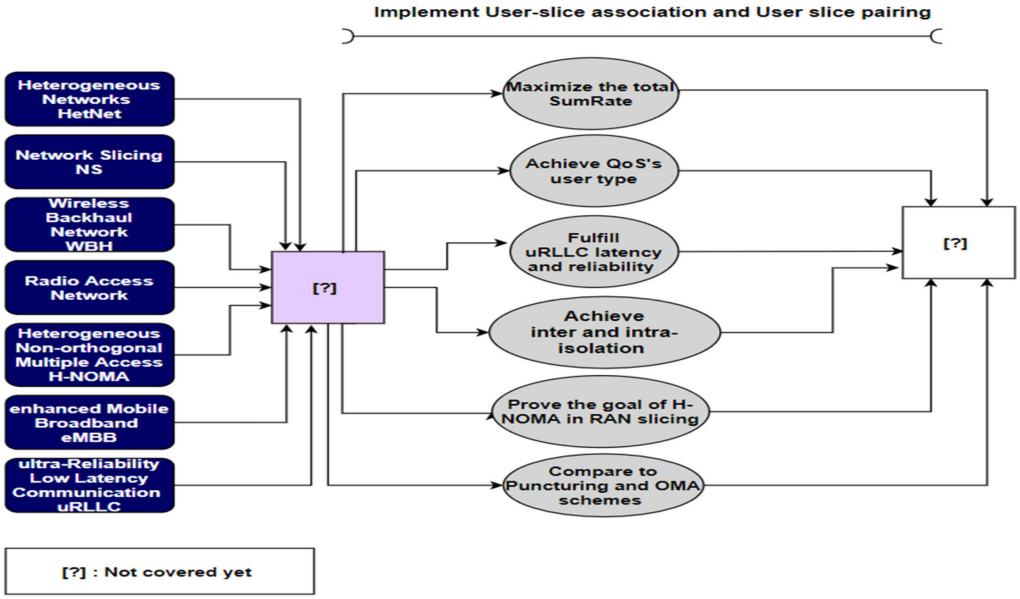

Figure 1. Contribution of this work (Table 1).

Table 1. Notation List.

Notions	Definitions	Notions	Definitions
$\mathscr{B}$	Set of base stations	R_u^{min}	Minimum DL required data rate for $u \in U$
$\mathscr{L}$	Set of uRLLC users	T_l^{max}	Maximum DL tolerant delay on the system for $l \in \mathscr{L}$
E	Set of eMBB users	$\mathscr{Z}_l(t)$	payload size for $l \in \mathscr{L}$
U	Set of eMBB and uRLLC users	λ_l	Arrival rate for $l \in \mathscr{L}$
$\mathscr{M}$	Number of mini slots in long TTI	C_{bl}	Block-length code for $l \in \mathscr{L}$, $b \in \mathcal{B}$
$\mathscr{T}$	Duration of a RAN time slot	ε	uRLLC reliability probability
δ	Duration of a RAN mini slot	$\mathscr{V}_{bl}^{nm}$	Channel dispersion for $l \in \mathscr{L}$, $b \in \mathcal{B}, n \in N_b^1, m \in \mathscr{M}$
$\mathcal{S}$	Set of slices per BS	T_{bl}^{WBH}	WBH mean transmission packet delay for $l \in \mathscr{L}$, $b \in \mathcal{B}$
I_s	Set of users in each slice s $\in \mathcal{S}$, BS $b \in \mathcal{B}$	T_{bl}^{RAN}	RAN mean transmission packet delay for $l \in \mathscr{L}$, $b \in \mathcal{B}$
N_b^s	Set of virtual PRBs per BS's slice $s \in \mathcal{S}, b \in \mathcal{B}$	SINR_{1b}	Received SINR for a SBS from MBS $b \in \mathcal{B}$, $b \neq 1$
$\mathscr{A}^U$	Binary UE-slice association matrix	η_{1b}	Channel gain between MBS and SBS $b \neq 1$
α_{bu}^{nm}	Power allocation coefficient for $u \in U$, $b \in \mathcal{B}$	o^2	Noise power
a_{bu}^{nm}	Association decision variable, for $u \in U$, $b \in \mathcal{B}, n \in N_b^s, m \in \mathscr{M}$	$SINR^{TH}$	Threshold SINR of WBH network
$\mathcal{X}^U$	Binary Superposition matrix by H-NOMA	br_{bu}	Number of bits for a single successful WBH transmission
$x_{u\grave{u}}^{nm}$	Superposition variable for $l \in \mathscr{L}$, $e \in E$ on $n \in N_b^s$, $m \in \mathscr{M}$	τ^{WBH}	WBH time slot
$\mathcal{Q}_{bu}^n$	Integer number of mini slots for $u \in U$, $b \in \mathcal{B}$	W^{WBH}	Bandwidth of the backhaul network
y_{bu}^{nm}	Received signal by user $u \in U$ from BS $b \in \mathcal{B}$ on $n \in N_b^s, m \in \mathscr{M}$	φ_{bl}	WBH required number of time slots $l \in \mathscr{L}$, $b \in \mathcal{B}$
$\mathscr{R}_{sb}^{rsv}$	Reserved rate for s $\in \mathcal{S}$, $b \in \mathcal{B}$	pr^{WBH}	WBH Transmission success probability of one packet in a single transmission
Γ_{bu}^{nm}	Received SINR for UE $u \in U$ from BS $b \in \mathcal{B}$ on $n \in N_b^s, m \in \mathscr{M}$	ϱ	Pathloss exponent of WBH link
ω_b^{nm}	Bandwidth of PRB $n \in N_b^s$ per $m \in \mathscr{M}$	T_{bl}^{tx}	RAN transmission time $l \in \mathscr{L}$, $b \in \mathcal{B}$
r_{bu}^{nm}	Data rate for UE $u \in U$ from BS $b \in \mathcal{B}$ on $n \in N_b^s$, per $m \in \mathscr{M}$	ρ_{bl}	Number of paired mini slots for $l \in \mathscr{L}$ with $e \in E$
R_{bu}	Overall DL received rate for $u \in U$ from BS $b \in \mathcal{B}$	D_l	maximum threshold RAN's packet delay
Ω_{bu}^{nm}	PRBs allocation decision variable, for $u \in U$, $b \in \mathcal{B}, m \in \mathscr{M}$	T_{bl}^{RTT}	HARQ retransmission delay $l \in \mathscr{L}$, $b \in \mathcal{B}$
γ_{bu}^m	Number of allocated PRBs for $u \in U$, $b \in \mathcal{B}, m \in \mathscr{M}$	T_{bl}^{WT}	Superimposed time for $l \in \mathscr{L}$ pairing with eMBB UE $e \in E$
f_{bu}^m	Number of minimum required PRBs for $u \in U$, $b \in \mathcal{B}, m \in \mathscr{M}$	$\overline{T}_{bl}$	Total mean transmission packet delay on the system for $l \in \mathscr{L}$, $b \in \mathcal{B}$

2. System Model

2.1. Network Model and Assumptions

We consider a two-tier downlink H-NOMA cellular network which consists of wireless backhaul and radio access networks with a single infrastructure provider (InP), where 1st tier consists of an MBS while 2nd tier models SBSs. The set of BSs is denoted as $b = \{1, 2, .., \ldots \mathscr{B}\}$ when $b = 1$ is a MBS. However, $b \neq 1$ is a SBS. We assume that the out-band operation mode is operated on the HetNet system. Thus, non-line-of-sight (NLOS) backhaul and access links are operated on different carriers; thus, separation in the frequency domain is provided to achieve efficient spectrum usage [43,44]. Also, we assume that all BSs utilize the same frequency band [45], resulting in SBSs being underlined by spectrum sharing with one MBS. Let $l = \{l_1, l_2, l_3, \ldots, \mathscr{L}\}$ and $e = \{e_1, e_2, e_3, \ldots, E\}$ denote the set of uRLLC and eMBB UEs, respectively, are randomly distributed over the entire network area. The set of all UEs containing uRLLC UEs and eMBB UEs is represented by $U = \mathscr{L} \cup E$, where the number of eMBB UEs is represented by $u \in U = \{1, 2, \ldots, k\}$ and the set of uRLLC UEs are indicated by $u = \{k+1, \ldots, U\}$. We assume the total transmitted power's MBS is equally distributed between WBH and RAN as P_{1b} where $b \epsilon B$, $b \neq 1$ and P_1^N,respectively. Each BS will have N_b physical resource blocks (PRBs) in RAN, distributed amongst its attached users depending on the available system band. We suppose that the overall transmitted power of each BS's RAN is allocated evenly over the available PRB (P_b^n). In this system design, we slice and assign PRBs as virtual resources into two RAN slices to serve uRLLC, and eMBB users, as illustrated in Figure 2. The set of all slices is denoted by $s \in \mathcal{S} = \{1, 2\}$, where $s = 1$ refers to the uRLLC slice and $s = 2$ is the eMBB slice. Each slice s has its own set of users denoted by I_s, let $\imath = \cup_s^{\mathcal{S}} \mathrm{I}_s$ where $| \cdot |$ is the cardinality of the set of all UEs. Considering non-orthogonal slicing [36], all virtual PRBs are used for uRLLC and eMBB transmissions at each slice's BS as $N_b^1 = N_b^2 = N_b \; \forall$ $b \in \mathcal{B}$, respectively. Since the connection and pairing processes are occurred on a larger time scale than the channel change, the small-scale fading term swings fast enough to be averaged in the measurements of its channel during the connection time.

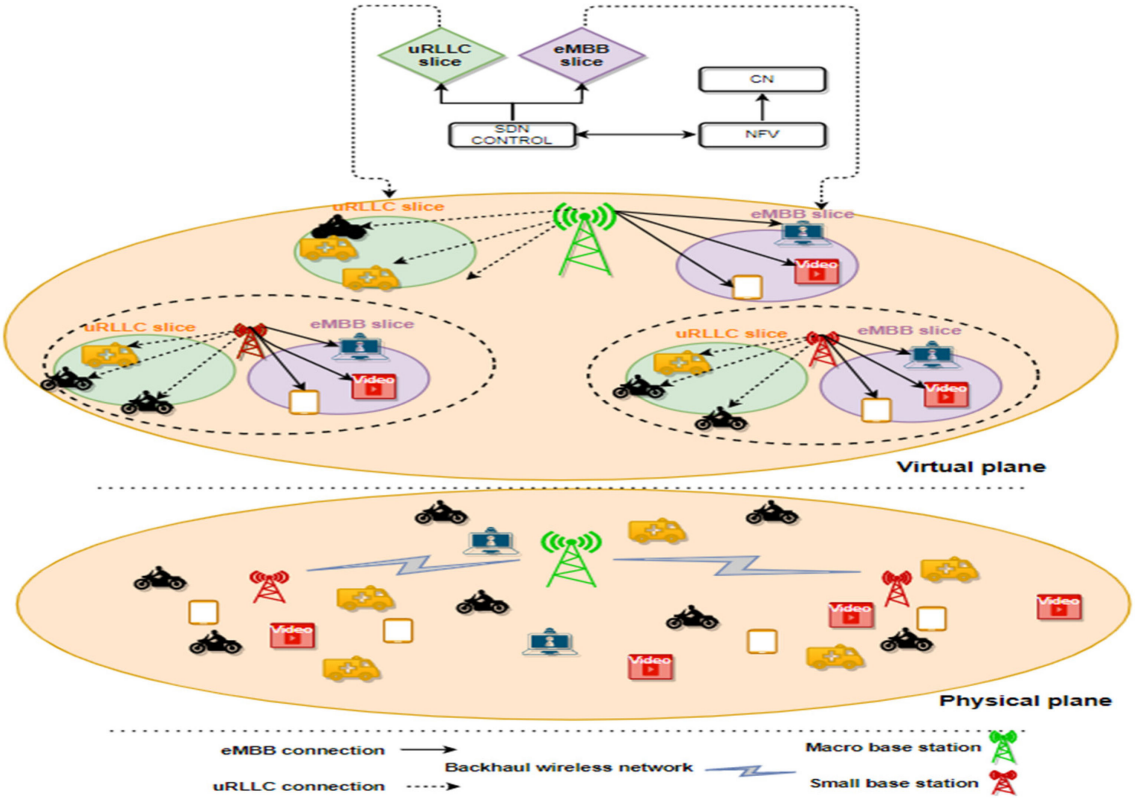

Figure 2. System Model.

2.2. Joint User Pairing and Association Matrix

Let us know a binary UE association matrix as $\mathscr{A}^U \in \{0,1\}^{\mathscr{B}\times U\times N^S_{\mathscr{B}}\times \mathscr{M}}$, where $a^{nm}_{bu} = 1$ if a UE u is only associated with one base station and assigned to a PRB n on mini-slot m, and $a^{nm}_{bu} = 0$, otherwise. Let us express a binary UE coupling matrix $\mathscr{X}^U \in \{0,1\}^{U\times U'\times N^S_{\mathscr{B}}\times \mathscr{M}}$, where $x^{nm}_{u\grave{u}} = 1$ if an eMBB UE u is paired with an uRLLC UE $\grave{u}$ by superposition scheme, at the same BS, and $x^{nm}_{u\grave{u}} = 0$, otherwise. $\mathscr{Q}^n_{bu}$ denotes the integer number of mini slots assigned to UE u when attached with BS b; if a UE is eMBB then $\mathscr{Q}^n_{bu} = \mathscr{M}$ (along TTI $\mathscr{T}$) as for a time slot; but if a UE is uRLLC so $\mathscr{Q}^n_{bu}$ is based on the paired eMBB UE.

To formulate the user association, we define the association matrix $\mathscr{A}^U$ as

$$a^{nm}_{bu} = \begin{cases} 1, & \text{if UE } u \text{ is associated with BS } b \\ 0, & \text{otherwise.} \end{cases} \tag{1}$$

The adoption of PD-NOMA scheme by the user u is represented by

$$x^{nm}_{u\grave{u}} = \begin{cases} 1, & \text{if user } u \text{ is superposited with } \grave{u} \text{are associated with the same BS ,} \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

2.3. Signalling Model

Two types of interferences should be considered in the proposed multi-cell H-NOMA system. Firstly, the inter-BS interference is generated by the signals transmitted from the un-associated BSs. Secondly, the intra-BS interference is produced by the co-channel interference of H-NOMA schemes. In any coalition, the received signal at user u is shown as follows [33,34] in Figure 3:

$$y^{nm}_{bu} = \underbrace{\sum_{h\in \mathfrak{B}\setminus\{b\}} \sum_{\grave{u}\in\{k+1,\ldots,U\}\setminus\{u\}} \sqrt{\alpha^{nm}_{hu}\, P^n_h}\, g^{nm}_{hu}\, x^{nm}_{u\grave{u}}\, r^{nm}_{hu}}_{\substack{\text{inter}-\text{BS interference} \\ \text{(Unwanted Signal)}}} + \underbrace{\sum_{b\in \mathfrak{B}} \sum_{u,\grave{u}\in U} \sqrt{(1-\alpha^{nm}_{bu}) P^n_b}\, g^{nm}_{bu}\, x^{nm}_{u\grave{u}}\, r^{nm}_{b\grave{u}}}_{\substack{\text{intra}-\text{BS interference} \\ \text{(Unwanted Signal: Remove using SIC)}}} + \underbrace{\sqrt{\alpha^{nm}_{bu}\, P^n_b}\, g^{nm}_{bu}\, x^{nm}_{u\grave{u}}\, r^{nm}_{bu}}_{\text{Wanted Signal}} + o^2 \tag{3}$$

where $\mathfrak{B}\setminus\{b\}$ is the collection of interfering nodes that utilise the same frequency channel, r^{nm}_{bu} the signal transmitted from the BS to the u UE, $\alpha^{nm}_{bu} \in [0, 1]$ denotes the power allocation coefficient for the UE u associated with BS b on PRB n; P^n_b denotes the transmit power of BS b over PRB n to UE u; g^{nm}_{bu} represents the channel gain between BS b and UE u at mini slot m and $(1-\alpha^{nm}_{bu})\, P^n_b\, g^{nm}_{bu}$ denote the received interference from the coexisting another user. Moreover, let $g_u = (g_{1u}, g_{2u}, \ldots, g_{\mathfrak{B}u}) \in G$ with the perfect CSI. o^2 is the additive noise power.

When H-NOMA is applied, each PRB can mostly serve two different services UEs. Implementing H-NOMA brings more sophisticated co-channel interferences (CCI) to the existing networks; to optimize the main problem without CCI employing perfect SIC, both eMBB and uRLLC UEs receive the superimposed signal, and each UE executes SIC to decode its message [46]. We assume the stronger user is a uRLLC UE which only can decode and remove CCI from the weak user is eMBB UE. While the eMBB UE with weaker channel conditions cannot wholly eliminate the interference of uRLLC UEs' signals, which results in maximized throughput for the eMBB UE with contrast [40] and prove it with the results in Section 5 Because our approach is based on uRLLC UE is superimposed with eMBB UE through a few mini slots and assigned high power for eMBB UE to reduce the latency by providing a high rate to the paired uRLLC UEs at a mini slot and achieving high reliability. Thus, the effect of CCI on eMBB UE is less than if the conventional pairing is along a TTI.

As well as, to guarantee inter-slice interference isolation between slices and ensure the QoS for each UE-slice according to H-NOMA technology, we assume $\mathscr{R}^{rsv}_{sb}$ is the reserved rate for each slice $s \in \mathscr{S} = \{1, 2\}$ in each BS [47,48]. The total rate of the cell is denoted as $\mathscr{R}^{rsv}_b = \mathscr{R}^{rsv}_{1b} + \mathscr{R}^{rsv}_{2b}$; based on the average downlink physical data rate of LTE system is achieved by [49].

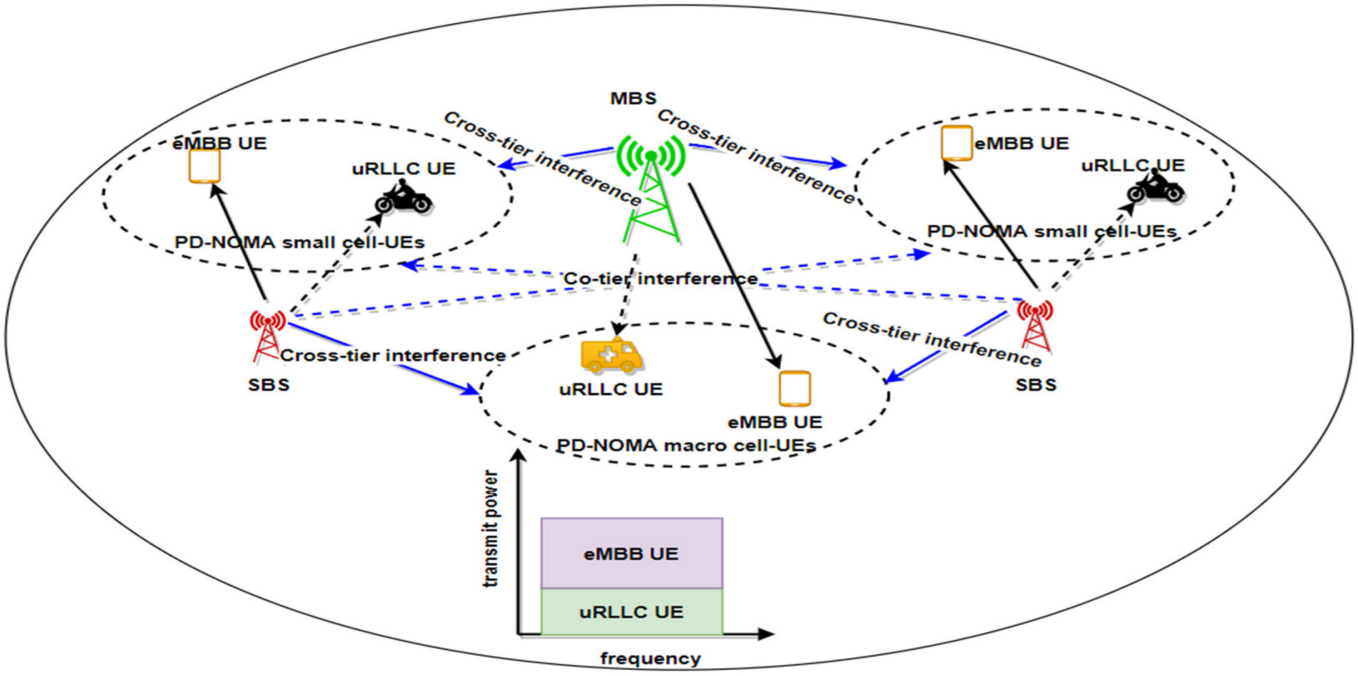

Figure 3. An illustration of eMBB and uRLLC users down-link PD-NOMA HetNet scheme.

Under the superposition scheme, the SINR Γ_{bu}^{nm} is received by the UE $u \in U$ from a BS $b \in \mathcal{B}$ on a PRB $n \in N_b^s$ per mini slot $m \in \mathcal{M}$ as, can be as follows

$$\Gamma_{bu}^{nm} = \frac{\alpha_{bu}^{nm} P_b^n g_{bu}^{nm}}{(1-\alpha_{bu}^{nm}) P_b^n g_{bu}^{nm} + \sum_{h \in \mathcal{B} \setminus \{b\}} \alpha_{hu}^{nm} P_{hu}^n g_{hu}^{nm} + o^2}, \tag{4}$$

2.4. Data Rate Expression Based on Shannon Capacity Model

The DL achievable data rate of a UE u attached with BS b with bandwidth ω_b^{nm} over one PRB n per mini slot m can be calculated as,

$$r_{bu}^{nm} = \omega_b^{nm} \log_2\left(1+\Gamma_{bu}^{nm}\right), \tag{5}$$

We denote the total received rate of a UE u, represented by R_{bu}, calculated as follows,

$$R_{bu} = \sum_{b \in \mathcal{B}} \sum_{u,\grave{u} \in U} a_{bu}^{nm} x_{u,\grave{u}}^{nm} \mathcal{Q}_{bu}^n \Omega_{bu}^{nm} \gamma_{bu}^m r_{bu}^{nm}, \tag{6}$$

Ω_{bu}^{nm} is the allocation indicator of the PRB n to UE u in BS b per mini slot m, where $\Omega_{bu}^{nm} = 1$ if a PRB n is allocated to a UE u and $\Omega_{bu}^{nm} = 0$, otherwise. Furthermore, γ_{bu}^m is the fraction of the PRB allocated to UE u and is calculated as $\gamma_{bu}^m = \left\lceil \frac{R_{bu}}{r_{bu}^{nm}} \right\rceil$, where $\lceil \cdot \rceil$ is the ceiling function.

$$\gamma_{bu}^m \geq f_{bu}^m, \tag{7}$$

When a UE u is associated with a BS b, γ_{bu}^m PRBs are allocated to a UE u, and this BS is duty to achieve the following condition

$$R_{bu} \geq R_u^{min}, \tag{8}$$

2.5. uRLLC Traffic

The uRLLC UEs are with delay-sensitive machine-type traffic and require high transmission reliability. Therefore, a UE l associates with a base station and superposes with an eMBB UE e who satisfy its required QoS constraint T_l^{max} & R_l^{min}.

2.5.1. uRLLC Data Rate Depending on Finite Block-Length Coding

The uRLLC's bursts of small payload sizes of $\mathcal{L}_l(t)$ bytes arrive at the network according to a Poisson Point Process (PPP) with an arrival rate of λ_l [payload/s] [50]. Due to the small packet size of uRLLC traffic, Shannon's data rate formulation cannot be used directly. Consequently, the DL data rate of uRLLC UE in (5) is edited based on [50] i.e.,

$$r_{bl}^{nm} = \omega_b^{nm} \log_2(1+\Gamma_{bl}^{nm}) - \sqrt{\frac{\mathcal{V}_{bl}^{nm}}{C_{bl}}}\, Q^{-1}(\varepsilon), \tag{9}$$

where C_{bl} is the block-length code, and $Q^{-1}(\varepsilon)$ is the inverse of the complementary Gaussian cumulative distribution Q-function with the probability of decoding error ε and $\mathcal{V}_{bl}^{nm}$ is the channel dispersion which is represented by

$$\mathcal{V}_{bl}^{nm} = 1 - \frac{1}{1+\left(1+\Gamma_{bl}^{nm}\right)}, \tag{10}$$

2.5.2. uRLLC Mean Packet Delay

In this system, the deadline interval of uRLLC UE l associated with BS b is contributed by the time processing of backhaul and RAN, are represented as T_{bl}^{WBH} and T_{bl}^{RAN}, respectively.

2.5.3. Wireless Backhaul Network

The analysis for backhaul connection behind the core gateway (GW) is neglected since it is common for all BSs and is assumed to be ideal. Thus, the received signal-to-interference-plus-noise ratio (SINR) to the b^{th} SBS from MBS on WBH is displayed as,

$$\mathrm{SINR}_{1b} = \frac{P_{1b}\,\eta_{1b}}{\sum_{h\in \mathrm{B}\setminus\{b\}} P_{1h}\,\eta_{1h} + o^2}, \tag{11}$$

where η_{1b} represents the channel gain between MBS and SBS. For the WBH network, the transmission succeeds if the received SINR is $\mathrm{SINR}_{1b} \leq SINR^{TH}$, $SINR^{TH}$ is a threshold SINR. Otherwise, the transmission has failed, necessitating retransmission. Clearly, the transmission success probability in a single transmission attempt is contingent upon the link length, bandwidth, and time slot length. The wireless backhaul transmission is time-slotted, and one packet is transmitted in each time slot. One hop is assumed for sub-6 GHz backhaul since the distance between the SBS and the MBS is often not great [51,52]. Meanwhile, the radio access network connects users with (macro cell or small cell) BSs through wireless links, which usually have only one hop. Considering the Shannon capacity formula, where the number of bits br_{bl} that can be transmitted from BS b to uRLLC UE l in a single successful transmission through backhaul time slot τ^{WBH} and W^{WBH} backhaul system bandwidth,

$$br_{bl} = \tau^{WBH}\, W^{WBH} \log_2(1+SINR^{TH}), \tag{12}$$

Thus, T_{bl}^{WBH} the mean transmission packet delay for uRLLC UE l over the sub-6 GHz WBH link, is based on $SINR_{1b}$ in (11), can be expressed as follows,

$$T_{bl}^{WBH} = \tau^{WBH}\, \varphi_{bl} \frac{1}{\rho r^{WBH}} = \tau^{WBH}\, \varphi_{bl}\, \frac{2}{\beta\, \sin(2\pi\beta)\, (SINR^{TH})^{2\beta}}, \tag{13}$$

where φ_{bl} is the required number of WBH time slots of a uRLLC UE for delivering its packet and equals $\left\lceil \frac{\mathcal{L}_l(\mathrm{t})}{br_{bl}} \right\rceil$; ρr^{WBH} is the transmission success probability in a single transmission attempt to calculate the average number of transfers required to deliver a packet successfully as $\frac{1}{\rho r^{WBH}}$, and $\beta = \frac{2}{\varrho}$, $\varrho > 2$ *is* pathloss exponent of WBH link.

2.5.4. Wireless RAN Network

T_{bl}^{RAN} consists of BS process time, UE process time, frame alignment time, queue time, and transmission time [31,50,53]. In this work, uRLLC and eMBB traffic are independent in slices' various services, so the queue of eMBB does not affect uRLLC. The system always has resources to service the uRLLC immediately as it arrives, which leads to the uRLLC UE having no queueing delay. BS and UE process time are bound by three OFDM symbol durations, which are very small and refer to the equipment computing capacity. As for, frame alignment time is upper bounded by the short TTI interval (δ), and transmission time T_{bl}^{tx} is related to the data rate of UE l. In superposition operation, UE l is paired with eMBB e, if its load R_l^{min} needs to be transmitted within the stipulated period δ, $T_{bl}^{tx} = \rho_{bl}\,\delta$, where ρ_{bl} is the number of mini slots in which uRLLC UE l pairs with eMBB UE e and the following condition is satisfied, $\max_{l \in \mathscr{L}} \frac{R_l^{min}}{R_{bl} T_{bl}^{tx}} \leq D_l$, where D_l indicates the maximum RAN packet delay threshold of uRLLC UE l and $R_{bl} = r_{bl}^{nm} * \gamma_{be}^{m}$. In case of failure, the packet is subject to additional retransmission delay T_{bl}^{RTT} is a round trip delay of a HARQ retransmission until either it is decoded successfully, or the maximum number of retransmissions is reached. This work considers uRLLC UE $u \in U$ scheduled with short TTI units composed of two OFDM symbols and Short HARQ RTT. Thus, we can determine the mean transmission packet delay T_{bl}^{RAN} of a uRLLC UE is as follows [50,53,54],

$$T_{bl}^{RAN} \cong T_{bl}^{tx} + T_{bl}^{WT} + T_{bl}^{RTT}, \tag{14}$$

where T_{bl}^{WT} is the superimposed mini-slot time for uRLLC UE l to pair with eMBB UE e. Each eMBB UE may be paired with more than one uRLLC UE, so the eMBB TTI is split into mini-slots; each paired uRLLC UE takes several ρ_{bl} mini-slots, and one mini slot is spaced to avoid intra-slice interference between uRLLC UEs that are paired.

To compute the average DL packet transmission delay of this system for uRLLC UE l [53,54], i.e.,

$$\overline{T}_{bl} = \begin{cases} T_{bl}^{RAN}, & \text{if } b = 1 \\ T_{bl}^{WBH} + T_{bl}^{RAN}, & \text{if } b \neq 1' \end{cases} \tag{15}$$

On the other side, the reliability of uRLLC can be improved by making sure that its outage probability is less than a certain threshold, according to [27],

$$Pr\left(\sum_{l \in \mathscr{L}} R_{bl} < a_{bl}^{nm} R_l^{min}\right) \leq \varepsilon, \tag{16}$$

If a uRLLC UE l connects with BS b and paires with an eMBB UE e, it is the duty of this BS and eMBB UE to achieve its QoS. i.e.,

$$R_{bl} \geq R_l^{min} \;\&\&\; \overline{T}_{bl} \leq T_l^{max}, \tag{17}$$

3. Problem Formulation

An association and pairing approach is now addressed by jointly considering eMBB and uRLLC UEs' different demands. Since the uRLLC UEs should operate on a constrained received latency, DL latency should be considered during the connection process for uRLLC users. In contrast to uRLLC UEs' needs, eMBB communications typically demand a high DL data rate. Thus, the DL data rate is a crucial association measure for eMBB users. Therefore, the UE-slice association and the UE-slice pairing indicators can be designed to maximize the overall DL data rate for eMBB UEs and minimize the DL transmitted latency for uRLLC UEs. Thus, the user-slice association (U − S. A) and the UE-slice pairing (U − S. P) problem can be represented as,

$$\text{O.P.T1} : \max_{\{a,x\}} [F_1, -F_2]^T, \tag{18}$$

It can be observed that the O.P.T1 (18) is a nonconvex mixed-integer programming problem, where the ideal solution is challenging to obtain. The subsequent subsection will convert the problem into a pure combinatorial optimization issue. Therefore, we adopt a suboptimal method and decompose the O.P.T1 into two designed sub-problems, as described in the next sub-sections. The first sub-problem is to create the user-slice association decision matrix $\mathscr{A}$. The second sub-problem is to create the user-slice pairing decision matrix $\mathscr{X}$. Thus, we suggest solving this problem using two-sided matching, as seen in the next section.

3.1. UE-Slice Association Sub-Problem

First, we concentrate on the association of users to BSs to boost overall DL data rates and minimize uRLLC latency while considering the RAN capacity for each BS and guaranteeing the intra-slice and inter-slice isolation. The U-S. A solution can be obtained by solving the following optimization sub-problem:

$$\text{O.P.T1} - \text{U} - \text{S. A} : \max_{\{a\}} \left[F_1(a), -F_2(a)\right]^T, \tag{19}$$

$$\text{S.t} \qquad (8), (16), (17), \tag{20}$$

$$\sum\nolimits_{u \in U} a_{bu}^{nm} \gamma_{bu}^{m} \leq N_b^s \quad \forall\, m \in \mathscr{M},\ s \in \mathcal{S},\ b \in \mathscr{B}, \tag{21}$$

$$\sum\nolimits_{u \in U} a_{bu}^{nm} R_{bu} \leq \mathscr{R}_{sb}^{rsv} \quad \forall\, m \in \mathscr{M},\ s \in \mathcal{S},\ b \in \mathscr{B} \tag{22}$$

$$\sum\nolimits_{,\ b \in \mathscr{B}} a_{bu}^{nm} = 1 \ \forall\, m \in \mathscr{M},\ s \in \mathcal{S},\ u \in U \tag{23}$$

$$\sum\nolimits_{,n \in N_b^s} \Omega_{bu}^{nm} \leq 1 \quad \forall\, m \in \mathscr{M},\ s \in \mathcal{S},\ u \in U,\ b \in \mathscr{B} \tag{24}$$

$$a_{bu}^{nm} \in \{0,1\} \quad \forall\, m \in \mathscr{M}, n \in N_b^s,\ s \in \mathcal{S},\ u \in U,\ b \in \mathscr{B} \tag{25}$$

$$f_{bu}^{m} \in \{1,2,\ldots,N_b^s\} \quad \forall\, m \in \mathscr{M},\ s \in \mathcal{S},\ u \in U,\ b \in \mathscr{B} \tag{26}$$

where $F_1(a) = \sum_{b \in \mathrm{B}} \sum_{u \in U} \sum_{s \in S} \sum_{n \in N_b^s} a_{bu}^{nm} R_{bu}$, that is, the overall data rate for eMBB and uRLLC UEs, and $F_2(a) = \sum_{b \in \mathscr{B}} \sum_{l \in \mathscr{L}} a_{bu}^{nm} \overline{T}_{bl}$, that is, the overall mean transmitted packet latency for uRLLC UEs. Note that to reduce the goal using the minus sign in (19), initially, we neglect the superimposed time $T_{bl}^{WT} = 0$. Constraint (21) allows the attached UE to utilize the PRBs without bypassing the base station resource harvest. Condition (22) ensures the inter-slice interference isolation in each BS. Constraint (23) ensures that each UE is associated with one BS or not attached at most, while condition (24) specifies that each PRB per TTI is allocated to one UE is in the same service's slice to achieve intra-slice isolation. Condition (25) indicates that the connection indicators have possible binary values, as for (26) indicates that it should occupy integer values and not exceed the maximum number of PRBs at each slice's BS.

Therefore, we observed that the O.P.T1 $-$ U $-$ S. A is a non-linear objective and complicated sub-problem due to the discrete features of the association indices. To solve the O.P.T1 $-$ U $-$ S. A sub-problem, we develop an algorithm for UE-slice association, in which the one-to-many matching game [36,37,39] is used, as described in Section 4.1.

3.2. UE-Slice Pairing Sub-Problem

Then, the optimum pairing for uRLLC UEs and eMBB UEs is achieved to completely maximize the network's total throughput and maintain the QoS requirements for users using the superposition (H-NOMA) scheme between eMBB and uRLLC UEs, additionally considering the intra-slice and inter-slice isolation and the limited capacity for the service's users, whereas we consider only the eMBB and uRLLC users are associated with the same

BS and consider the same objective of (19) with fixed a^t, $\forall\ \mathcal{T}$. Therefore, the (U − S. P) sub-problem can be designed as follows:

$$\text{O.P.T1} - \text{U} - \text{S. P} : \max_{\{x\}} [F_1(x), -F_2(x)]^T, \tag{27}$$

$$\text{S.t} \qquad (8), (16), (17) \tag{28}$$

$$x_{u\grave{u}}^{nm} \in \{0,1\} \quad \forall\ m \in M,\ n \in N_b^s,\ s \in \mathcal{S},\ u \in \mathcal{L}, \grave{u} \in E,\ b \in \mathcal{B} \tag{29}$$

$$T_{bl}^{WT} \in \{1,2,\ldots,\mathcal{M}\} \quad \forall\ l \in \mathcal{L},\ b \in \mathcal{B} \tag{30}$$

where $F_1(x) = \sum_{b \in \text{B}} \sum_{u,\grave{u} \in U} \sum_{s \in S} \sum_{n \in N_b^s} x_{u\grave{u}}^{nm}\ R_{bu}$, and $F_2(x) = \sum_{b \in \mathcal{B}} \sum_{l \in \mathcal{L}} x_{u\grave{u}}^{nm}\ \overline{T}_{bl}$. Condition (28) means that the pairing indicators values are binary values. In this case, constraint (29) should take into consideration the superimposed time of pairing in our calculations, which is a random integer mini-slot time. Therefore, a one-to-one matching game [40,55,55] is used to acquire the UE-slice pairing solution, which is explained in Section 4.2.

4. Solution Using Matching Game-Based U-S. A and U-S. P Algorithms

In this section, we propose two matching algorithms to solve the U-S. A and U-S. P formulated sub-problems. Since O.P.T1 − U − S. A and O.P.T1 − U − S. P sub-problems are formulated as a one-to-many and one-to-one matching game, respectively. The matching function of U − S. A is defined by $\mu_{\text{U}-\text{S. A}}$ with a tuple $(\mathcal{B}, U, \Theta, \succ_{\mathcal{B}}, \succ_U)$ and the U − S. P is represented as $\mu_{\text{U}-\text{S. P}}$ with a tuple $\left(\mathcal{L}, E, \widetilde{\Theta}, \succ_{\mathcal{L}}, \succ_E\right)$. In the U − S. A matching function, $\succ_{\mathcal{B}} = \{\succ_b\}_{b \in \mathcal{B}}$ and $\succ_U = \{\succ_u\}_{u \in U}$ indicates the preference relations of the BSs and UEs players, respectively. As for the U − S. P matching function, $\succ_{\mathcal{L}} = \{\succ_l\}_{l \in \mathcal{L}}$ and $\succ_E = \{\succ_e\}_{e \in E}$ denote the preference relations of the associated uRLLC UEs and eMBB UEs with the same BS, respectively. Assume that Θ and $\widetilde{\Theta}$ are the quota for O.P.T1 − U − S. A and O.P.T1 − U − S. P, respectively.

4.1. One-Sided Matching Based Solution Approach Sub-Problem (19)

There are diverse preferences utilities that can be supplied for players depending on their tendencies. For O.P.T1 − U − S. A, we erect the preference relations of UEs (eMBB and uRLLC UEs) and BSs for selecting the best match according to the utility functions listed as $\Psi_u(b)$ and $\Psi_b(u)$.

4.1.1. UEs Utility Function

Since eMBB UEs are data-rate-hungry, we utilize the DL attainable data rate as the utility function $\Psi_e(b)$, which can be expressed as follows:

$$\Psi_e(b) = R_{be}, \tag{31}$$

Meanwhile, uRLLC UEs are deemed urgent-latency UEs; thus, the utility attained by an uRLLC UE when it is associated with the BS is introduced as a function of the DL transmitted latency and reliability as follows:

$$\Psi_l^1(b) = R_{bl}\ \&\ \Psi_l^2(b) = \overline{T}_{bl}, \tag{32}$$

4.1.2. Base Stations Utility Function

The base station's utility function must be efficient for the user-slice association. Where each BS wants to serve UEs with the highest rate to optimize the overall DL throughput based on the following utility function:

$$\Psi_b(u) = R_{bu}, \tag{33}$$

The major details of the $U-S.$ A algorithm are depicted in Algorithm 1. After initialization, each UE erects its preference relations $H_u \succ_u$ based on (31) and (32) (step 4). Similarly, the preference list for each BS H_b is constructed based on the preference utility (23). At each iteration It, all unassigned UEs (eMBB and uRLLC UEs) send attachment requests to their preferred BSs (steps 5–9). Each BS will then determine whether to accept the requests or not based on its defined utility and quota Θ_b (Θ_b^{rem} and Θ_b^{th} available radio resources and rate for each slice of a BS) that restricts the number of attached UEs to avert Quality of Experience (QoE) retrogression. If a BS has sufficient Θ_b^{rem} and Θ_b^{th} to accept, it accepts the request and updates Θ_b^{rem}, Θ_b^{th} and $\mu_{U-S.\ A}^{(It)}(b)$ (steps 10–14). Otherwise, if the quota is not sufficient, but the utility (31)/(32) of the requesting UE u is more significant than a UE $\hat{u}$ that was accepted in the previous round, the BS b determines all its current matches $\hat{u}$ which have a worse ranking than according to steps H_b^{It} (steps 15–17). Each least preferred $\hat{u}$ is then sequentially removed, and Θ_b^{rem}, Θ_b^{th}, $\mu_{U-S.\ A}^{(It)}(b)$ and the minimum preference list (ξ_{mp}) is updated until it can be accepted or there is no more to reject (steps 18–29). After rejecting all UEs $\hat{u}$ and the available PRBs at BS are still insufficient to admit a UE u. Thus, it is refused and assigned to ξ_{mp} (steps 30–31), i.e., the refused UEs will send an attachment request in the next round to the next BS in their preference lists. Consequently, BS b removes ξ_{mp} its less preferred UEs from the respective lists H_b. Similarly, these UEs remove a BS b from their preference lists (lines 32–33). After many iterations It, there are no more bidding UEs; thus, the algorithm converges to a stable match.

4.2. One-Sided Matching Based Solution Approach Sub-Problem (26)

We can better understand the philosophy of the $U-S.$ P sub-problem and the solution scheme with an interpretative example, as shown in Figure 4. After the association approach, we assume that each BS has sufficient traffic for the connected eMBB and uRLLC UEs. Thus, eMBB users are scheduled on all available PRBs at the beginning of a time slot, and each eMBB UE owns n PRBs long TTI and fixed through the slot. When uRLLC traffic comes abruptly to its associated BS b (in any mini slot of the current slot), the BS b tries to superimpose a uRLLC user with a suitable eMBB user guaranteeing all constraints of each slice and UEs through the mini slot m. Then, the scheduler immediately aims to schedule such traffic with ρ_{bl} paired mini slots. Due to the hard latency bindings of uRLLC traffic, we apply the H-NOMA mechanism to be applied for the uRLLC traffic in this paper. Generally, uRLLC traffic has a small payload size and, thus, requires a fraction of all mini slots for such traffic. However, the questions are about selecting the suitable eMBB and the number of paired mini slots that eMBB users currently occupy are the best to be superimposed, keeping the objective of the sub-problem (26) in mind. Therefore, BS balances the data rate of eMBB users in each long TTI, ultimately achieving the goal of (26) on a long-run basis. In the superimposing scenario of eMBB and uRLLC users associated with the same BS, each uRLLC UE is only able to couple with one eMBB UE per mini slot m, but eMBB UE may be paired with a maximum of $\widetilde{\Theta}$ uRLLC UEs per time slot t. Consequently, each uRLLC UE is initially assigned to an eMBB UE. Once this uRLLC UE has been paired with the appropriate eMBB UE, it is deleted, and the procedure is repeated with the remaining set of uRLLC and eMBB UEs. Therefore, we assume the time slot t as two sub-time slots and apply a one-to-one matching game between eMBB and uRLLC users. As for $O.P.T1-U-S.$ P, we erect the preference lists of eMBB and uRLLC UEs based on the following utility functions as $\Lambda_e(l)$ and $\Lambda_l(e)$.

Algorithm 1: Matching Game for User-Slice Association

Input: U, $\mathscr{B}$
Output: Find stable Matching $\mu^{(It)}_{U-S.\ A}$
Initialization: $It = 0$, $\mu^{(It)}_{U-S.\ A} = \left\{\mu^{(It)}_{U-S.\ A}(b),\ \mu^{(It)}_{U-S.\ A}(u)\right\}_{\forall u \in U, b \in \mathscr{B}} = \varnothing$, $\xi_{mp} = \varnothing$, $z^{(It)}_b = \varnothing$, $\Theta_b = \left\{\Theta^{rem}_b = N^s_b,\ \Theta^{th}_b = 0\right\}_{s \in \mathscr{S},\, b \in \mathscr{B}}$
Every UE construct $\succ_u$ using $\Psi_u(b)$
Repeat
  $It \leftarrow It + 1$
  for $b \in \mathscr{B}$ **do**
    for $u \in U$ do with b as its best preferred in H^{It}_u **do**
      while $u \notin \mu^{(It)}_{U-S.\ A}(b)$ and $H^{It}_u \neq \varnothing$ **do**
        If $N^s_b \geq f^m_{bu}$ and $\mathscr{R}^{rsv}_{sb} \geq R_{bu}$ **then**
          If $\Theta^{rem}_b \geq f^m_{bu}$ and $\Theta^{th}_b < \mathscr{R}^{rsv}_{sb}$ **then**
            $\mu^{(It)}_{U-S.\ A}(b) \leftarrow \mu^{(It)}_{U-S.\ A}(b) \cup \{u\}$,
            $\Theta^{rem}_b \leftarrow \Theta^{rem}_b - f^m_{bu}$,
            $\Theta^{th}_b \leftarrow \Theta^{th}_b + R_{bu}$
          else
            $\hat{H}^{It}_b \leftarrow \left\{\hat{u} \in \mu^{(It)}_{U-S.\ A}(b) \middle| u \succ_b \hat{u}\right\}$,
            $\xi_{mp} \leftarrow \hat{u} \in \hat{H}^{It}_b$
            while $\hat{H}^{It}_b \cup (\Theta^{rem}_b < f^m_{bu})$ **do**
              $\mu^{(It)}_{U-S.\ A}(b) \leftarrow \mu^{(It)}_{U-S.\ A}(b) \backslash \{\xi_{mp}\}$,
              $\hat{H}^{It}_b \leftarrow \hat{H}^{It}_b \backslash \{\xi_{mp}\}$,
              $\Theta^{rem}_b \leftarrow \Theta^{rem}_b - f^m_{bu}$
              $\Theta^{th}_b \leftarrow \Theta^{th}_b + R_{bu}$,
              $\xi_{mp} \leftarrow \hat{u} \in \hat{H}^{It}_b$
              If $\Theta^{rem}_b \geq f^m_{bu}$ and $\Theta^{th}_b < \mathscr{R}^{rsv}_{sb}$ **then**
                $\mu^{(It)}_{U-S.\ A}(b) \leftarrow \mu^{(It)}_{U-S.\ A}(b) \cup \{u\}$,
                $\Theta^{rem}_b \leftarrow \Theta^{rem}_b - f^m_{bu}$,
                $\Theta^{th}_b \leftarrow \Theta^{th}_b + R_{bu}$
              else
                $\xi_{mp} \leftarrow u$,
                $z^{(It)}_b = \{z \in H^{It}_b | \ \xi_{mp} \succ_b z\} \cup \{\xi_{mp}\}$
                for $z \in z^{(It)}_b$ **do**
                  $H^{It}_z \in H^{It}_z \backslash \{b\}$, $H^{It}_b \leftarrow H^{It}_b \backslash \{z\}$,
  until $\mu^{(It)}_{U-S.\ A}(b) = \mu^{(It-1)}_{U-S.\ A}(b)$

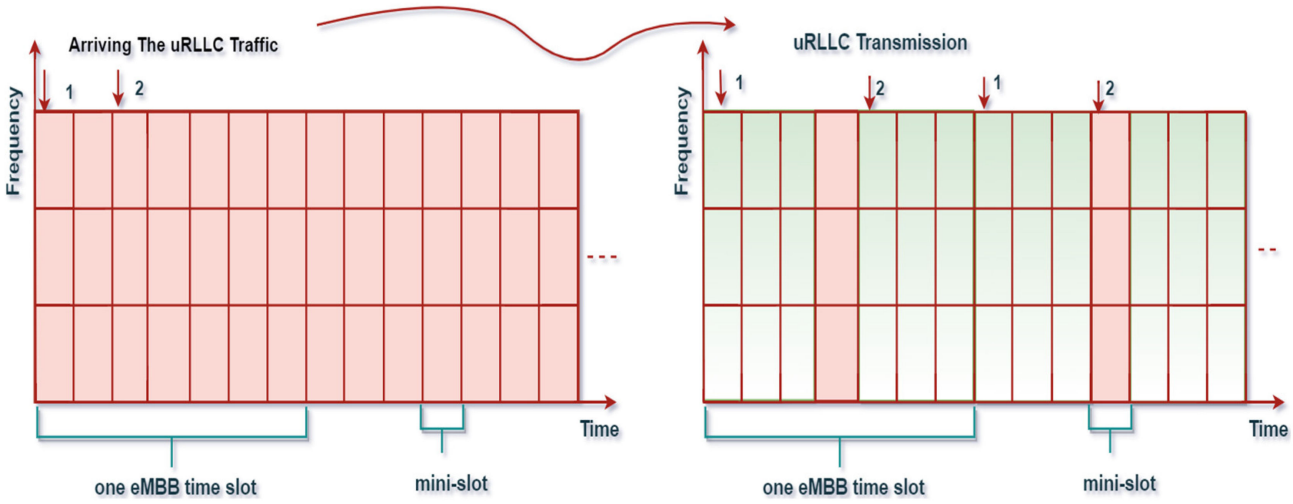

Figure 4. Illustration of the proposed heterogeneous- NS(HNS) pairing.

4.2.1. eMBB UEs Utility Function

Firstly, each eMBB UE selects the lowest rate's uRLLC UE for matching due to owning the entire PRB without a partner. Its preference utility can be represented as:

$$\Lambda_e(l) = R_{bl}, \tag{34}$$

4.2.2. uRLLC UEs Utility Function

On the other hand, each uRLLC UE ranks attached eMBB UEs at the same BS with the shortest transmission delay and the highest rate. Thus, each uRLLC UE's preference utility is dependent on the following:

$$\Lambda_l(e) = R_{be}, \tag{35}$$

Algorithm 2 describes the suggested matching-based U − S. P algorithm using the same technique as Algorithm 1. So, as UEs and BS have fixed preference relationships, algorithms are the deferred acceptance algorithm (DAA) for two-sided matching that leads to a stable match [37,42]. We estimate the computational complexity of the described algorithms using the significant *O* notation. Even under the worst-case scenario, the proposed algorithm's ongoing complexity may be recognized and compared to other algorithms on both approaches, as shown in Table 2.

Algorithm 2: Matching Game for User-Slice Pairing

Input: $E_b, \mathscr{L}_b, \mathscr{B}$
Output: Find stable Matching $\mu^{(It)}_{U-S.\ P}$
Initialization: $It = 0$, $\mu^{(It)}_{U-S.\ P} = \left\{\mu^{(It)}_{U-S.\ P}(l_b),\ \mu^{(It)}_{U-S.\ P}(e_b)\right\}_{l_b\in\mathscr{L}, e_b\in E, b\in\mathscr{B}} = \varnothing$, $\xi_{mp} = \varnothing$, $z^{(It)}_{l_b} = \varnothing$, $\widetilde{\Theta} = \left\{\widetilde{\Theta}^{rem}_{l_b} = 0, \widetilde{\Theta}^{rem}_{e_b} = 0, \widetilde{\Theta}^{mod}_{e_b} = 0\right\}_{e_b\in E, l_b\in\mathscr{L},\ b\in\mathscr{B}}$
Repeat
  $It \leftarrow It + 1$
  for $b \in \mathscr{B}$ **do**
    Sorts E_b attached with b in descending order based on the achieved rate and b construct $\succ_{l_b}$ using $\Lambda_{l_b}(e_b)$
    for $e_b \in E_{\text{sort}}$ **do**
      $H^{It}_{e_b} \leftarrow$ BS b construct $\succ_{e_b}$ using $\Lambda_{e_b}(l_b)$
      $\widetilde{\Theta}^{rem}_{e_b} \leftarrow \frac{\mathscr{L}_b}{E_b}$
      $\widetilde{\Theta}^{mod}_{e_b} \leftarrow |\mathscr{L}_b|\ mod\ |E_b|$
      If $\widetilde{\Theta}^{mod}_{e_b} > 0$ **do**
        $\widetilde{\Theta}^{rem}_{e_b} \leftarrow \widetilde{\Theta}^{rem}_{e_b} + 1$
      while $e_b \notin \mu^{(It)}_{U-S.\ P}(l_b)$ and $H^{It}_{e_b} \neq \varnothing$ **do**
        If $\Theta^{rem}_{l_b} = 0$ **then**
          $\mu^{(It)}_{U-S.\ P}(l_b) \leftarrow \mu^{(It)}_{U-S.\ P}(l_b) \cup \{e_b\}$,
          $\widetilde{\Theta}^{rem}_{e_b} \leftarrow \widetilde{\Theta}^{rem}_{e_b} - 1$,
          $\Theta^{rem}_{l_b} \leftarrow \Theta^{rem}_{l_b} + 1$
        else
          $\hat{H}^{It}_{l_b} \leftarrow \left\{\hat{e_b} \in \mu^{(It)}_{U-S.\ P}(e_b) \middle| e_b \succ_{l_b} \hat{e_b}\right\}$,
          $\xi_{mp} \leftarrow \hat{e_b} \in \hat{H}^{It}_{l_b}$
          while $\hat{H}^{It}_{l_b} \cup (\widetilde{\Theta}^{rem}_{l_b} = 1)$ **do**
            $\Lambda^{(It)}_{DP}(l_b) \leftarrow \Lambda^{(It)}_{DP}(l_b) \backslash \{\xi_{mp}\}$,
            $\hat{H}^{It}_{l_b} \leftarrow \hat{H}^{It}_{l_b} \backslash \{\xi_{mp}\}$,
            $\Theta^{rem}_{\hat{e_b}} \leftarrow \Theta^{rem}_{\hat{e_b}} + 1$
            $\Theta^{rem}_{l_b} \leftarrow \Theta^{rem}_{l_b} - 1$
            $\xi_{mp} \leftarrow \hat{e_b} \in \hat{H}^{It}_{l_b}$
            $\Lambda^{(It)}_{DP}(l_b) \leftarrow \Lambda^{(It)}_{DP}(l_b) \cup \{e_b\}$,
until $\mu^{(It)}_{U-S.\ P}(l_b) = \mu^{(It-1)}_{U-S.\ P}(l_b)$

Table 2. Computational complexity of user-slice association, user-slice pairing schemes.

U-S. A Scheme	Complexity	U-S. P Scheme	Complexity
DAA [37,39,55]	$O(U\mathscr{B})$	DAA	$O(\mathscr{L}E)$
EAA [56]	$O(U\mathscr{B}\log(U\mathscr{B}))$	EAA	$O(\mathscr{L}E\log(\mathscr{L}E))$
GA [57]	$O(U^2\mathscr{B}^2)$	GA	$O(\mathscr{L}^2E^2)$
MAX-SINR	$O(U^2\mathscr{B}^2\log(U\mathscr{B}))$	MAX-SINR	$O(\mathscr{L}E\log(E))$

5. Results

5.1. Simulation Setup

We consider a two-tier HetNet with four SBSs deployed within a macro cell coverage region radius of 250 m to assess the proposed matching game-based UE-slice association and UE-slice pairing algorithms. SBSs and UEs are uniformly distributed in the coverage area of the MBS. The simulations are dependent on 1000 runs, and the results are averaged. The essential simulation parameters are given in Table 3 [15,54].

Table 3. Simulation Parameters.

Parameter	Value
Transmit power of macro-BS	46 dBm
Transmit power of pico-BS	33 dBm
Backhaul bandwidth	40 MHZ
Backhaul timeslot	25 µs
5GRAN bandwidth	20 MHz
Number of PRBs	100
Inter-site distance	500 m
Pathloss between MBS and device	$128.1 + 37.6 * \log(d)$
Pathloss between SBS and device	$140.7 + 36.7 * \log(d)$
eMBB-rate threshold	$[1 - 4]$ Mbps
uRLLC rate threshold	$[0.1 - 1.6]$ Mbps
Modulation	4-QAM, 64 QAM for uRLLC and eMBB, respectively
uRLLC packet size	$[32 - 256]$ bytes
PHY numerology	15 kHz subcarrier spacing;12 subcarriers per PRB;2-OFDM symbols TTI (0.143 ms)
HARQ	Asynchronous HARQ with chase combining, and 4 TTI round trip time; Max 6 HARQ retransmissions.

5.2. Simulation Results

Most research projects on the coexistence between eMBB and uRLLC concur that the ratio of uRLLC to eMBB UEs is double per cell [15,54,58]. So, we assign the quota of U-S. P algorithm as $\widetilde{\Theta} = 2$. Thus, the paired uRLLC UE superimposes with an eMBB for $\rho_{bl} = 3$ mini slots.

The number of eMBB UEs and uRLLC UEs are 25 and 50, respectively [15,54]. The minimum required rate levels of eMBB and uRLLC UEs are 4 and 0.5 Mbps, respectively. The submitted algorithm's effectiveness is compared with the Early Acceptance algorithm (EAA)-based matching game, Max-SINR, and Greedy algorithm (GA). All these algorithms are displayed in both cases of SINR $\Gamma_{be}^{nm} > \Gamma_{bl}^{nm}$ and $\Gamma_{bl}^{nm} > \Gamma_{be}^{nm}$, for different techniques (H-NOMA, OMA & Puncturing).

We show the total DL throughput of uRLLC UEs with the different numbers of uRLLC when $\Gamma_{be}^{nm} > \Gamma_{bl}^{nm}$ and $\Gamma_{bl}^{nm} > \Gamma_{be}^{nm}$, in Figure 5 respectively. When uRLLC UEs grow, the uRLLC overall throughput rises in all schemes. In UE-slice association and pairing operations, uRLLC UEs are more interested in latency and reliability than DL data rate. We observe how similar the two shapes are, where our submitted H-NOMA algorithms are still superior in performance over DAA- PUNC and OMA strategies. To explain it, a uRLLC UE is paired with an eMBB UE and occupies the eMBB's PRBs per ρ_{bl} mini slots. As for using the puncturing technique, an uRLLC UE is overlapped with an eMBB UE and occupies an eMBB's PRBs per one mini slot. Meanwhile using the OMA technique, each uRLLC UE takes up one PRB per long TTI.

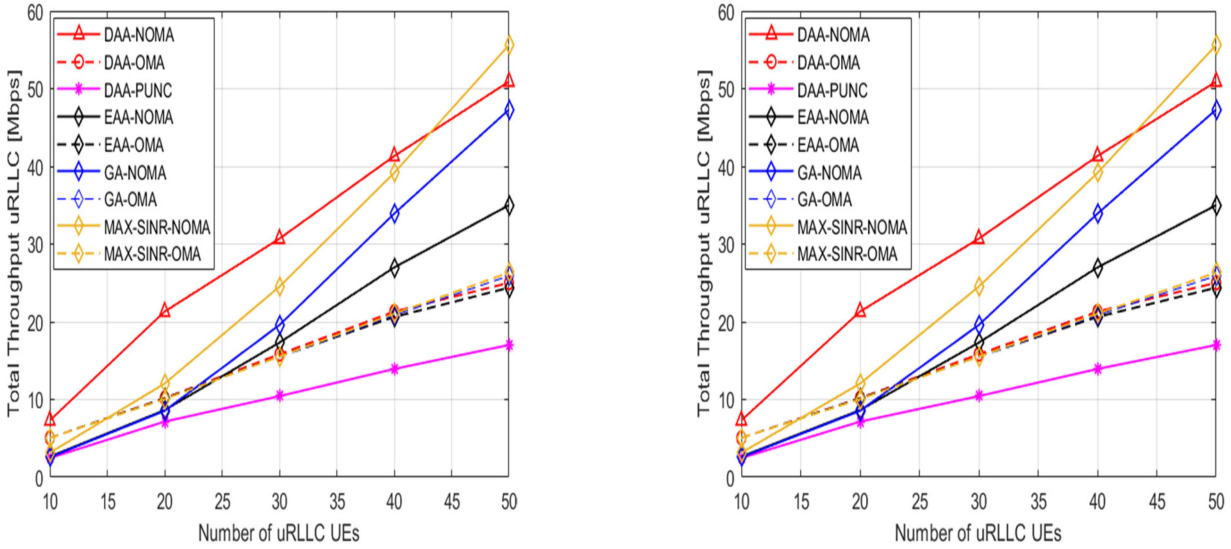

Figure 5. The DL total throughput for uRLLC UEs versus number of uRLLC UEs. (**a**) $\Gamma_{be}^{nm} > \Gamma_{bl}^{nm}$. (**b**) $\Gamma_{bl}^{nm} > \Gamma_{be}^{nm}$.

After documenting our idea to distribute power among various users' PRB, it is time to display the other outputs and meet the goals of this study; as shown in Figure 6, the DL average throughput of eMBB UEs is against the number of uRLLC UEs. The average throughput of eMBB UEs falls as the number of uRLLC UEs grows. Even though the DL data rate is not an appealing attachment and pairing metric for uRLLC UEs, it suggests that UEs are vying to associate and pair with the best elected BSs based on their preference lists. The proposed algorithm and EAA-NOMA consider the QoS requirements for eMBB UEs during the association process. Thus, the minimum required DL data rate for eMBB UEs can be achieved, as seen in Figure 6. However, the suggested approach, which incorporates the QoS requirements for eMBB UEs throughout the association and pairing operations, has the lowest rate of loss for eMBB service and outperforms EAA-NOMA, particularly at many uRLLC UEs. This happens because EAA-NOMA does not achieve the best optimum solution since it allows the highest preference list to associate and pair until the system capacity is complete, but the eMBB UEs' minimum wanted DL data rate is assured; thus, not achieving the spirit of cooperation, which is based on a matching game. Without considering eMBB UE rate needs and the available PRBs at each BS, the performance of both Max-SINR and GA is worse than the proposed algorithm. In Max-SINR, most of the users are associated with MBS.

Meanwhile, SBSs serve a relatively small number of users. Accordingly, many MBS-attached users may not be served because of the limited number of resources available in the MBS. In GA, BSs are unconcerned with the UEs' QoS requirements and focus only on achieving a higher overall DL rate system. These are explanations for why MAX-SNR and GA have the worst performance rate compared to other schemes.

Figure 7 illustrates various techniques' outage probabilities as the number of uRLLC UEs grows. The outage probability is seen in the light as the likelihood that a uRLLC UE consumes latency processes is more than the maximum DL tolerant delay. It can be observed that, at a high density of uRLLC UEs, the outage probability of our submitted scheme stays very low compared to that of PUNC and OMA technology if applied with any algorithm. Simulation results show promising outage probability achievement for uRLLC traffic at various schemes, as shown in Table 4.

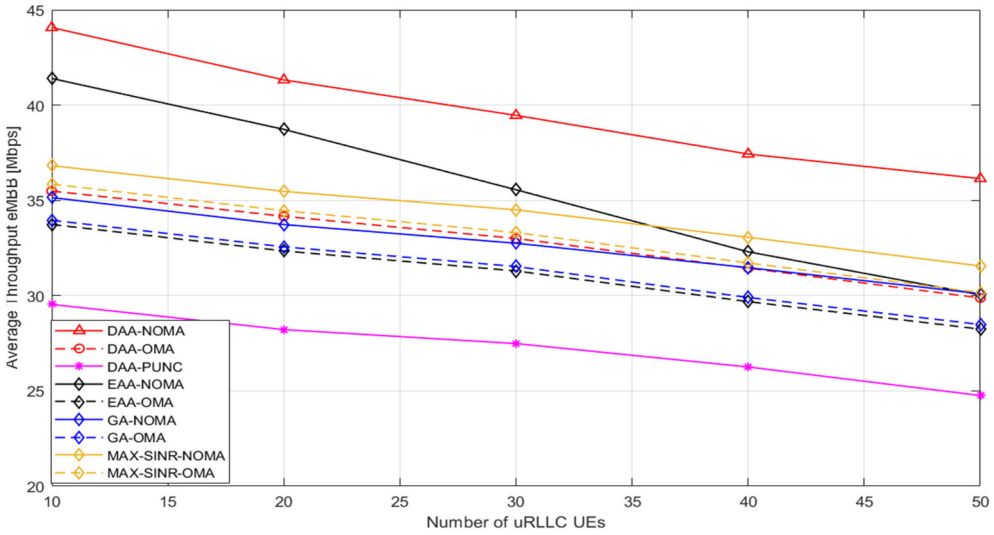

Figure 6. The DL average throughput for eMBB UEs versus number of uRLLC UEs.

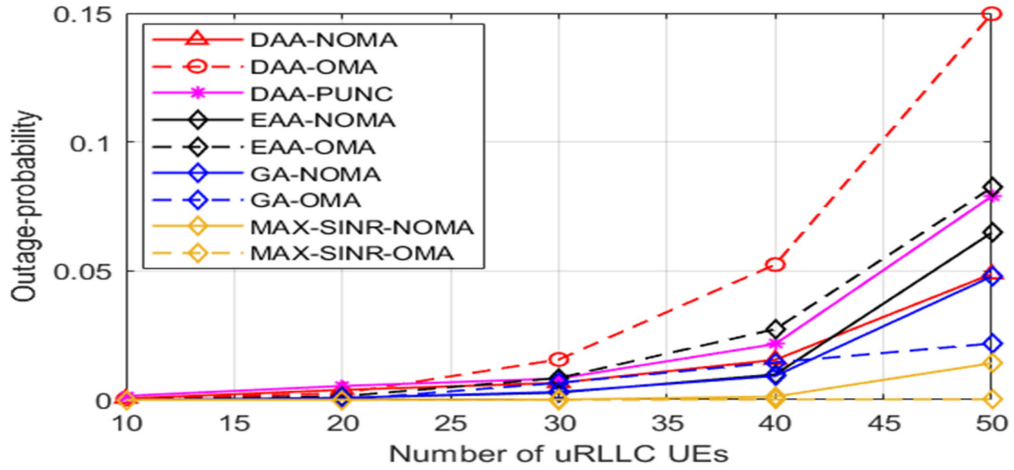

Figure 7. The outage-probability for uRLLC UEs versus number of uRLLC UEs.

Table 4. Outage Probability for uRLLC Traffic.

Schemes	Outage Probability
DAA-NOMA	99.04%
DAA-OMA	97.03%
DAA- PUNC	98.3%
EAA-NOMA	98.85%
EAA-OMA	98.64%
GHA-NOMA	99.03%
GHA-OMA	99.42%
MAX-SINR-NOMA	99.78%
MAX-SINR-OMA	99.99%

Figure 8 illustrates the overall DL data rate for eMBB UEs when the number of eMBB UEs grows from 10 to 50 and the number of uRLLC UEs = 50 for two cases $\Gamma_{be}^{nm} > \Gamma_{bl}^{nm}$ and $\Gamma_{bl}^{nm} > \Gamma_{be}^{nm}$, respectively. It can be observed that the total sum rate for eMBB UEs increases if the number of eMBB UEs increases in both cases, but the proposed H-NOMA algorithm demonstrates higher rates if $\Gamma_{be}^{nm} > \Gamma_{bl}^{nm}$ (due to its use of higher power and suffering from slight CCI), as shown in Figure 8. Furthermore, the suggested correlation's performance approach outperforms all other schemes, which affirms the effectiveness of the proposed

attachment and pairing approaches because our suggested algorithm considers the type of user, its demands, and intra and inter-isolation constraints. The total sum rate by OMA and Puncturing techniques is lower than that of H-NOMA, indicating that H-NOMA is more sum rate efficient than OMA. The explanations are that, as for the OMA technique, we assume that the transmitted power of an OMA-PRB is $\frac{1}{2}$ the power of a NOMA-PRB; thus, intra and inter-isolation is achieved using the OMA technique. So, each eMBB slice has a fraction of $\frac{N_b^2}{N_b} = \frac{4}{5}$ based on [59]; In the Puncturing technique, when uRLLC UEs are overlapped with eMBB UE, it is scheduled on all eMBB's PRBs per one mini slot. Simulation results confirm promising gains of up to 40% DL sum rate improvement for eMBB traffic compared to "OMA" and "Puncturing" schemes, as presented in Table 5. So, our assumed framework achieves the highest overall DL data rate for both different users, as illustrated in Figures 5 and 8.

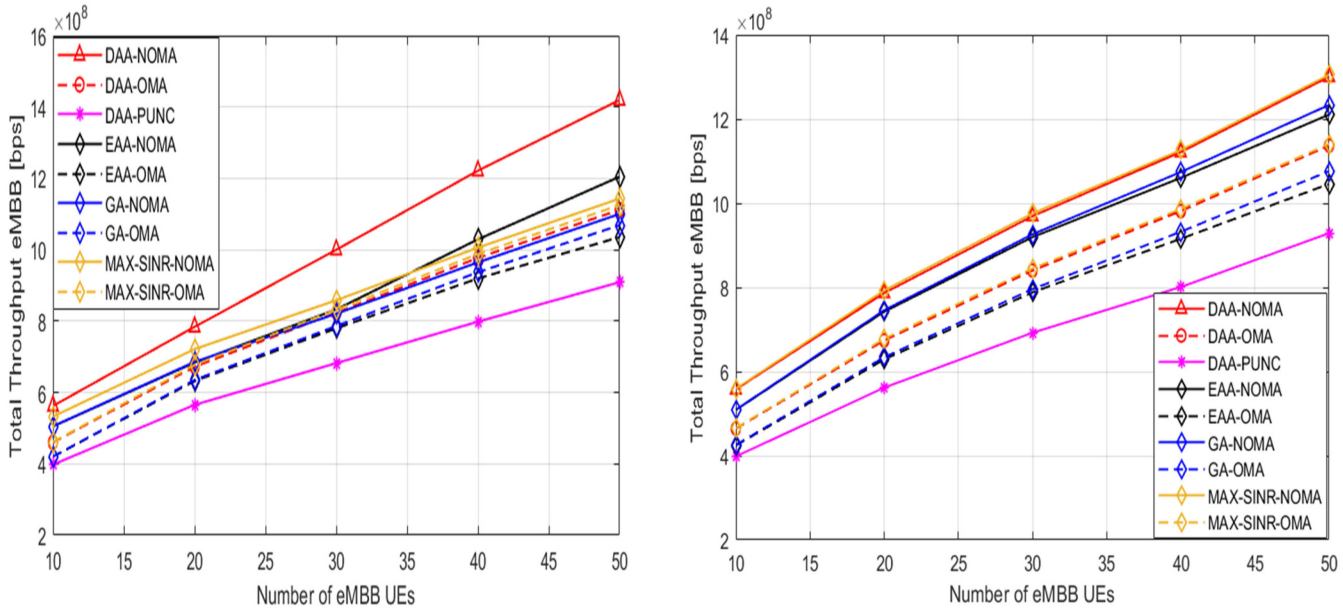

Figure 8. The DL total throughput for eMBB UEs versus number of eMBB UEs. (**a**) $\Gamma_{be}^{nm} > \Gamma_{bl}^{nm}$. (**b**) $\Gamma_{bl}^{nm} > \Gamma_{be}^{nm}$.

Table 5. DL Sum Rate gains for eMBB Traffic.

Compared Algorithms	DL Sum Rate Gains
DAA-OMA	24.3%
DAA- PUNC	41.9%
EAA-NOMA	16.4%
EAA-OMA	34.3%
GHA-NOMA	18.3%
GHA-OMA	32.3%
MAX-SINR-NOMA	14%
MAX-SINR-OMA	23.6%

Figure 9 depicts the outage likelihood of the various approaches estimated as the number of eMBB UEs grows. The outage probability of an eMBB UE is represented as not achieving the minimum required DL rate. It can be observed that, at a high density of users, the outage probability of our suggested scheme stays very low compared to other strategies. Simulation results show promising outage probability improvements for eMBB traffic at various schemes, as shown in Table 6.

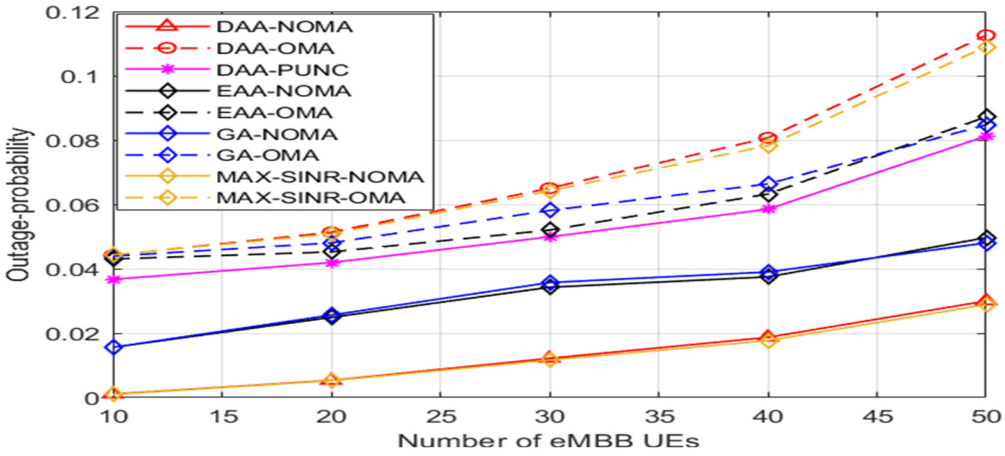

Figure 9. The outage-probability for eMBB UEs versus number of eMBB UEs.

Table 6. Outage Probability for eMBB Traffic.

Schemes	Outage Probability
DAA-NOMA	98.6%
DAA-OMA	92.9%
DAA-PUNC	94.6%
EAA-NOMA	96.7%
EAA-OMA	94%
GHA-NOMA	96.7%
GHA-OMA	93.9%
MAX-SINR-NOMA	98.6%
MAX-SINR-OMA	93%

Figure 10 indicates the DL latency per uRLLC as the minimum DL packet size for uRLLC UE increases. This figure proves that the proposed algorithm performs better than the other approaches. Our proposed algorithm proves that each uRLLC packet can be sent in less than 1 msec, regardless of whether it is associated with MBS or SBS, based on the transmission time (backhaul and RAN transmission time), waiting pairing time, and HARQ time. This figure shows that the proposed H-NOMA approach to all algorithms performs better than DAA-PUNC and OMA. The simulation illustrates promising gains of up to 95% and 80% in latency improvement for uRLLC traffic compared to "OMA" and "Puncturing" schemes, respectively, as shown in Table 7.

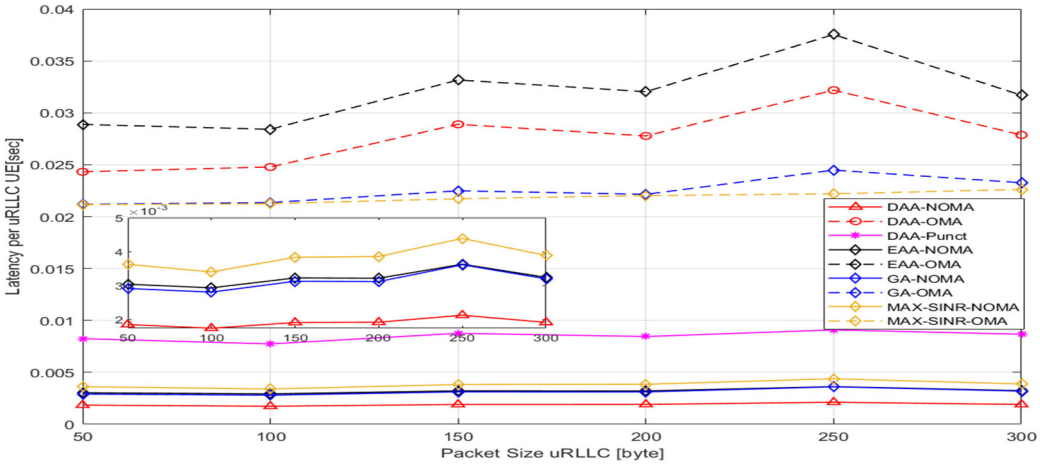

Figure 10. The DL average throughput for eMBB UEs versus offered load uRLLC.

Table 7. DL latency gains for uRLLC Traffic.

Compared Algorithms	DL Latency Gains
DAA-OMA	92%
DAA- PUNC	77%
EAA-NOMA	40%
EAA-OMA	93%
GHA-NOMA	38%
GHA-OMA	91%
MAX-SINR-NOMA	49%
MAX-SINR-OMA	91%

6. Discussion

In the future, further research can be pursued to investigate the following open issues. First, we can use one of the proposed ways to prove its effectiveness in minimizing co-channel interference, such as (a) a hybrid overlay-underlay spectrum access scheme in heterogeneous networks to improve energy efficiency [60]. (b) cooperative relaying networks with NOMA-HetNet [61]. A relay uses a decode-and-forward (DF) scheme to improve the performance of the far user in terms of the outage and throughput. (c) a downlink underlay cognitive radio-NOMA (CR-NOMA) HetNet to reduce CCI and enhance the performance of spectrum sharing [55]. Second, we can combine HetNets with aerially controlled networks such as Unmanned Aerial Vehicles (UAVs) are required to overcome these challenges. UAV-aided HetNet shows lower transmission delay (lower latency) and better average jitter compared to without UAV-based HetNet, because UAVs control the transmission of packets with efficient utilization of available bandwidth from source to destination [62]. Third, we can use FANETs [63], in which multiple UAVs cooperate and establish an ad hoc network in a multi-UAV scenario to enhance network performance. Finally, we will leverage potential deep learning approaches to provide coexisting radio resources for various services. Using a learning-based strategy for flexible joint UE association and pairing will prevent coexistence concerns in 5G and wireless networks.

7. Conclusions

In this letter, joint user-slice association and user-slice pairing algorithms for eMBB and uRLLC UEs in HetNet are suggested based on game theory. The suggested algorithms are designed as an optimization problem to maximize the overall DL throughput while assuring the slice's QoS, minimizing the latency of uRLLC traffic, and achieving intra and inter-isolation using flexible rate isolation. To improve its tractability, we have divided the problem between (1) dynamic association matching of BSs and UEs and (2) superimposing of uRLLC and eMBB UEs. Matching game algorithms for U-S. A and U-S. P are proposed to solve these subproblems. Moreover, we propose a dynamic eMBB and uRLLC clustering technique called H-NOMA in HetNet systems to balance system performance. The computation complexity for the proposed algorithms is analyzed. Simulation results have confirmed that the submitted algorithm achieved throughput significantly above the compared algorithms and defeated uRLLC UE latency deterioration. The proposed algorithm had the highest throughput and the lowest latency compared with DAA-OMA, DAA-PUNC, EAA-NOMA, GA-NOMA, GA-OMA, MAX-SINR-NOMA, and MAX-SINR-OMA.

Author Contributions: Conceptualization, M.A.R.; Data curation, M.A.R.; Formal analysis, M.A.R. and A.A.E.-H.; Methodology, M.A.R.; Validation, M.A.R., O.E.-G. and A.A.E.-H.; Writing—original draft, A.A.E.-H.; Writing—review & editing, M.A.R. and A.A.E.-H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

RAN	Radio access network
HetNet	heterogeneous networks
NS	Network slicing
QoS	Quality of service
uRLLC	Ultra-reliability low latency communications
eMBB	Enhanced mobile broadband
UE	User equipment
NOMA	Non-orthogonal multiple access
SIC	Successive interference cancellation
U-S. A	UE-slice association
U-S. P	UE-slice Pairing
MBS	Macro base station
SBS	Small bas station
WBH	Wireless backhaul network
H-NOMA	Heterogenous non-orthogonal multiple access
ITU	International Telecommunications Union
mMTC	massive Machine Type Communication
HARQ	Hybrid Automatic Repeat reQuest
CN	Core network
SDN	Software-Defined Networking
NFV	Network Functions Virtualization
EDF	Earliest deadline first
RSMA	Rate-splitting multiple access
CSI	Channel status information
MIMO	multiple-input multiple-output
RIS	Reconfigurable intelligent surface
EPS	Enhanced Pre-emptive Scheduling
C-RAN	Cloud radio access network
DRL	Deep reinforcement learning
MC	Multi-connectivity
PtrNet	Pointer network
RAT	Radio access technology
OFDM	Orthogonal frequency-division multiple access
TTI	Transmission time interval
InP	Infrastructure provider
NLOS	Non-line-of-sight
PRB	Physical resource block
CCI	Co-channel interferences
QoE	Quality of experience
DAA	Deferred acceptance algorithm
EAA	Early acceptance algorithm
MAX-SINR	Maximum-signal-to-interference-plus-noise ratio
GA	Greedy algorithm
OMA	Orthogonal multiple access
PUNC	Puncturing
C-NOMA	Cooperative non-orthogonal multiple access
CR-NOMA	Cognitive radio non-orthogonal multiple access
FANET	Flying Ad Hoc Networks

References

1. Parkvall, S.; Dahlman, E.; Furuskar, A.; Frenne, M. NR: The New 5G Radio Access Technology. *IEEE Commun. Stand. Mag.* **2017**, *1*, 24–30. [CrossRef]
2. Zhang, G.; Quek, T.Q.S.; Kountouris, M.; Huang, A.; Shan, H. Fundamentals of heterogeneous backhaul design—Analysis and optimization. *IEEE Trans. Commun.* **2016**, *64*, 876–889. [CrossRef]
3. Celik, A.; Tsai, M.-C.; Radaydeh, R.M.; Al-Qahtani, F.S.; Alouini, M.-S. Distributed Cluster Formation and Power-Bandwidth Allocation for Imperfect NOMA in DL-HetNets. *IEEE Trans. Commun.* **2018**, *67*, 1677–1692. [CrossRef]
4. Islam, S.M.R.; Avazov, N.; Dobre, O.A.; Kwak, K.-S. Power-Domain Non-Orthogonal Multiple Access (NOMA) in 5G Systems: Potentials and Challenges. *IEEE Commun. Surv. Tutor.* **2016**, *19*, 721–742. [CrossRef]
5. Zhang, X.; Haenggi, M. The Performance of Successive Interference Cancellation in Random Wireless Networks. *IEEE Trans. Inf. Theory* **2014**, *60*, 6368–6388. [CrossRef]
6. Series, M. *Minimum Requirements Related to Technical Performance for IMT-2020 Radio Interface (s)*; Report 2410-0; ITU: Geneva, Switzerland, 2017.
7. Kazmi, S.A.; Khan, L.U.; Tran, N.H.; Hong, C.S. *Network Slicing for 5G and Beyond Networks*; Springer: Berlin/Heidelberg, Germany, 2019; Volume 1.
8. Li, X.; Samaka, M.; Chan, H.A.; Bhamare, D.; Gupta, L.; Guo, C.; Jain, R. Network Slicing for 5G: Challenges and Opportunities. *IEEE Internet Comput.* **2017**, *21*, 20–27. [CrossRef]
9. Sattar, D.; Matrawy, A. Optimal slice allocation in 5G core networks. *IEEE Netw. Lett.* **2019**, *1*, 48–51. [CrossRef]
10. Barakabitze, A.A.; Ahmad, A.; Mijumbi, R.; Hines, A. 5G network slicing using SDN and NFV: A survey of taxonomy, architectures and future challenges. *Comput. Netw.* **2020**, *167*, 106984. [CrossRef]
11. Elayoubi, S.E.; Jemaa, S.B.; Altman, Z.; Galindo-Serrano, A. 5G RAN slicing for verticals: Enablers and challenges. *IEEE Commun. Mag.* **2019**, *57*, 28–34.
12. Carugi, M. Key features and requirements of 5G/IMT-2020 networks. In *ITU Arab Forum on Emerging Technologies*; Internetsociety.org: Reston, VA, USA, 2018.
13. Abedin, S.F.; Alam MG, R.; Kazmi, S.A.; Tran, N.H.; Niyato, D.; Hong, C.S. Resource allocation for ultra-reliable and enhanced mobile broadband IoT applications in fog network. *IEEE Trans. Commun.* **2018**, *67*, 489–502.
14. Liu, D.; Wang, L.; Chen, Y.; Elkashlan, M.; Wong, K.-K.; Schober, R.; Hanzo, L. User Association in 5G Networks: A Survey and an Outlook. *IEEE Commun. Surv. Tutor.* **2016**, *18*, 1018–1044. [CrossRef]
15. Guo, T.; Suárez, A. Enabling 5G RAN slicing with EDF slice scheduling. *IEEE Trans. Veh. Technol.* **2019**, *68*, 2865–2877. [CrossRef]
16. Popovski, P.; Trillingsgaard, K.F.; Simeone, O.; Durisi, G. 5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view. *Ieee Access* **2018**, *6*, 55765–55779.
17. Ding, Z.; Lei, X.; Karagiannidis, G.K.; Schober, R.; Yuan, J.; Bhargava, V.K. A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends. *IEEE J. Sel. Areas Commun.* **2017**, *35*, 2181–2195.
18. Aldababsa, M.; Toka, M.; Gökçeli, S.; Kurt, G.K.; Kucur, O. A tutorial on nonorthogonal multiple access for 5G and beyond. *Wirel. Commun. Mob. Comput.* **2018**, *2018*, 9713450.
19. Liu, Y.; Qin, Z.; Elkashlan, M.; Ding, Z.; Nallanathan, A.; Hanzo, L. Non-orthogonal multiple access for 5G and beyond. *arXiv* **2018**, arXiv:1808.00277.
20. Wang, Y.; Ren, B.; Sun, S.; Kang, S.; Yue, X. Analysis of non-orthogonal multiple access for 5G. *China Commun.* **2016**, *13*, 52–66.
21. Dos Santos, E.J.; Souza, R.D.; Rebelatto, J.L. Rate-Splitting Multiple Access for URLLC Uplink in Physical Layer Network Slicing With eMBB. *IEEE Access* **2021**, *9*, 163178–163187. [CrossRef]
22. Chen, Q.; Wang, J.; Jiang, H. URLLC and eMBB Coexistence in MIMO Non-orthogonal Multiple Access Systems. *arXiv* **2021**, arXiv:2109.05725.
23. Almekhlafi, M.; Arfaoui, M.A.; Elhattab, M.; Assi, C.; Ghrayeb, A. Joint Resource Allocation and Phase Shift Optimization for RIS-Aided eMBB/URLLC Traffic Multiplexing. *IEEE Trans. Commun.* **2022**, *70*, 1304–1319. [CrossRef]
24. Esswie, A.A.; Pedersen, K.I. Opportunistic Spatial Preemptive Scheduling for URLLC and eMBB Coexistence in Multi-User 5G Networks. *IEEE Access* **2018**, *6*, 38451–38463. [CrossRef]
25. Matera, A.; Kassab, R.; Simeone, O.; Spagnolini, U. Non-orthogonal eMBB-URLLC radio access for cloud radio access networks with analog fronthauling. *Entropy* **2018**, *20*, 661.
26. Alsenwi, M.; Tran, N.H.; Bennis, M.; Bairagi, A.K.; Hong, C.S. eMBB-URLLC resource slicing: A risk-sensitive approach. *IEEE Commun. Lett.* **2019**, *23*, 740–743.
27. Alsenwi, M.; Tran, N.H.; Bennis, M.; Pandey, S.R.; Bairagi, A.K.; Hong, C.S. Intelligent resource slicing for eMBB and URLLC coexistence in 5G and beyond: A deep reinforcement learning based approach. *IEEE Trans. Wirel. Commun.* **2021**, *20*, 4585–4600. [CrossRef]
28. Anand, A.; De Veciana, G.; Shakkottai, S. Joint scheduling of URLLC and eMBB traffic in 5G wireless networks. *IEEE/ACM Trans. Netw.* **2020**, *28*, 477–490.
29. Bairagi, A.K.; Munir, S.; Alsenwi, M.; Tran, N.H.; Alshamrani, S.S.; Masud, M.; Han, Z.; Hong, C.S. Coexistence mechanism between eMBB and uRLLC in 5G wireless networks. *IEEE Trans. Commun.* **2020**, *69*, 1736–1749. [CrossRef]
30. Tebe, P.I.; Ntiamoah-Sarpong, K.; Tian, W.; Li, J.; Huang, Y.; Wen, G. Using 5G network slicing and non-orthogonal multiple access to transmit medical data in a mobile hospital system. *IEEE Access* **2020**, *8*, 189163–189178.

31. Zhang, K.; Xu, X.; Zhang, J.; Zhang, B.; Tao, X.; Zhang, Y. Dynamic multiconnectivity based joint scheduling of eMBB and uRLLC in 5G networks. *IEEE Syst. J.* **2020**, *15*, 1333–1343. [CrossRef]
32. Wang, K.; Liu, Y.; Ding, Z.; Nallanathan, A.; Peng, M. User Association and Power Allocation for Multi-Cell Non-Orthogonal Multiple Access Networks. *IEEE Trans. Wirel. Commun.* **2019**, *18*, 5284–5298. [CrossRef]
33. Hasan, N.; Rizvi, S.; Shabbir, A. A Clustered PD-NOMA in an Ultra-Dense Heterogeneous Network with Improved System Capacity and Throughput. *Appl. Sci.* **2022**, *12*, 5206. [CrossRef]
34. Wang, K.; Liu, Y.; Ding, Z.; Nallanathan, A. User Association in Non-Orthogonal Multiple Access Networks. In Proceedings of the 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, USA, 20–24 May 2018; pp. 1–6. [CrossRef]
35. Yu, Z.; Hou, J. Research on Interference Coordination Optimization Strategy for User Fairness in NOMA Heterogeneous Networks. *Electronics* **2022**, *11*, 1700. [CrossRef]
36. Amine, M.; Kobbane, A.; Ben-Othman, J. New network slicing scheme for UE association solution in 5G ultra dense HetNets. In Proceedings of the ICC 2020–2020 IEEE International Conference on Communications (ICC), Dublin, Ireland, 7–11 June 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6.
37. Le TH, T.; Tran, N.H.; LeAnh, T.; Hong, C.S. User matching game in virtualized 5G cellular networks. In Proceedings of the 2016 18th Asia-Pacific Network Operations and Management Symposium (APNOMS), Kanazawa, Japan, 5–7 October 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1–4.
38. Ma, M.; Wong, V.W. Joint user pairing and association for multicell NOMA: A pointer network-based approach. In Proceedings of the 2020 IEEE International Conference on Communications Workshops (ICC Workshops), Dublin, Ireland, 7–11 June 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6.
39. Gu, Y.; Saad, W.; Bennis, M.; Debbah, M.; Han, Z. Matching theory for future wireless networks: Fundamentals and applications. *IEEE Commun. Mag.* **2015**, *53*, 52–59. [CrossRef]
40. Manzoor, A.; Kazmi, S.A.; Pandey, S.R.; Hong, C.S. Contract-based scheduling of URLLC packets in incumbent EMBB traffic. *IEEE Access* **2020**, *8*, 167516–167526. [CrossRef]
41. Elhattab, M.K.; Elmesalawy, M.M.; Salem, F.M.; Ibrahim, I.I. Device-aware cell association in heterogeneous cellular networks: A matching game approach. *IEEE Trans. Green Commun. Netw.* **2018**, *3*, 57–66. [CrossRef]
42. Anany, M.; Elmesalawy, M.M.; Abd El-Haleem, A.M. Matching game-based cell association in multi-rat HetNet considering device requirements. *IEEE Internet Things J.* **2019**, *6*, 9774–9782.
43. Gora, J.; Redana, S. In-band and out-band relaying configurations for dual-carrier LTE-advanced system. In Proceedings of the 2011 IEEE 22nd International Symposium on Personal, Indoor and Mobile Radio Communications, Toronto, ON, Canada, 11–14 September 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 1820–1824.
44. Sharma, A.; Ganti, R.K.; Milleth, J.K. Joint backhaul-access analysis of full duplex self-backhauling heterogeneous networks. *IEEE Trans. Wirel. Commun.* **2017**, *16*, 1727–1740. [CrossRef]
45. Peng, M.; Zhang, K.; Jiang, J.; Wang, J.; Wang, W. Energy-Efficient Resource Assignment and Power Allocation in Heterogeneous Cloud Radio Access Networks. *IEEE Trans. Veh. Technol.* **2014**, *64*, 5275–5287. [CrossRef]
46. Dos Santos, E.J.; Souza, R.D.; Rebelatto, J.L.; Alves, H. Network slicing for URLLC and eMBB with max-matching diversity channel allocation. *IEEE Commun. Lett.* **2019**, *24*, 658–661. [CrossRef]
47. Parsaeefard, S.; Dawadi, R.; Derakhshani, M.; Le-Ngoc, T. Joint User-Association and Resource-Allocation in Virtualized Wireless Networks. *IEEE Access* **2016**, *4*, 2738–2750. [CrossRef]
48. Rezvani, S.; Yamchi, N.M.; Javan, M.R.; Jorswieck, E.A. Resource allocation in virtualized CoMP-NOMA HetNets: Multi-connectivity for joint transmission. *IEEE Trans. Commun.* **2021**, *69*, 4172–4185.
49. Poornima, P.; Laxminarayana, G.; Rao, D.S. Performance analysis of channel capacity and throughput of lte downlink system. *Int. J. Comput. Netw. Commun.* **2017**, *9*, 55–69.
50. Karimi, A.; Pedersen, K.I.; Mahmood, N.H.; Pocovi, G.; Mogensen, P. Efficient Low Complexity Packet Scheduling Algorithm for Mixed URLLC and eMBB Traffic in 5G. In Proceedings of the 2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring), Kuala Lumpur, Malaysia, 28 April–1 May 2019; pp. 1–6. [CrossRef]
51. Zhang, G.; Quek, T.Q.; Huang, A.; Kountouris, M.; Shan, H. Backhaul-aware base station association in two-tier heterogeneous cellular networks. In Proceedings of the 2015 IEEE 16th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), Stockholm, Sweden, 28 June–1 July 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 390–394.
52. Maaz, D.; Galindo-Serrano, A.; Elayoubi, S.E. URLLC user plane latency performance in new radio. In Proceedings of the 2018 25th International Conference on Telecommunications (ICT), Saint-Malo, France, 26–28 June 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 225–229.
53. Elsayed, M.; Erol-Kantarci, M. AI-enabled radio resource allocation in 5G for URLLC and eMBB users. In Proceedings of the 2019 IEEE 2nd 5G World Forum (5GWF), Dresden, Germany, 30 September–2 October 2019; IEEE: Piscataway, NJ, USA; pp. 590–595.
54. Kazmi, S.M.A.; Tran, N.H.; Saad, W.; Han, Z.; Ho, T.M.; Oo, T.Z.; Hong, C.S. Mode Selection and Resource Allocation in Device-to-Device Communications: A Matching Game Approach. *IEEE Trans. Mob. Comput.* **2017**, *16*, 3126–3141. [CrossRef]
55. Liang, W.; Ding, Z.; Li, Y.; Song, L. User Pairing for Downlink Non-Orthogonal Multiple Access Networks Using Matching Algorithm. *IEEE Trans. Commun.* **2017**, *65*, 5319–5332. [CrossRef]
56. Alizadeh, A.; Vu, M. Distributed User Association in B5G Networks Using Early Acceptance Matching Game. *IEEE Trans. Wirel. Commun.* **2020**, *20*, 2428–2441. [CrossRef]

57. Zalghout, M.; Helard, J.-F.; Crussiere, M.; Abdul-Nabi, S.; Khalil, A. A Greedy Heuristic Algorithm for Context-Aware User Association and Resource Allocation in Heterogeneous Wireless Networks. In Proceedings of the 2017 IEEE 86th Vehicular Technology Conference (VTC-Fall), Toronto, ON, Canada, 24–27 September 2017; pp. 1–7. [CrossRef]
58. Pradhan, A.; Das, S. Joint Preference Metric for Efficient Resource Allocation in Co-Existence of eMBB and URLLC. In Proceedings of the 2020 International Conference on COMmunication Systems & NETworkS (COMSNETS), Bengaluru, India, 7–11 January 2020; pp. 897–899. [CrossRef]
59. Ginige, N.U.; Manosha, K.S.; Rajatheva, N.; Latva-aho, M. Admission control in 5G networks for the coexistence of eMBB-URLLC users. In Proceedings of the 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), Antwerp, Belgium, 25–28 May 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–6.
60. Darabi, M.; Jamali, V.; Lampe, L.; Schober, R. Hybrid Puncturing and Superposition Scheme for Joint Scheduling of URLLC and eMBB Traffic. *IEEE Commun. Lett.* **2022**, *26*, 1081–1085. [CrossRef]
61. Yue, X.; Liu, Y.; Kang, S.; Nallanathan, A.; Ding, Z. Outage performance of full/half-duplex user relaying in NOMA systems. In Proceedings of the 2017 IEEE International Conference on Communications (ICC), Paris, France, 21–25 May 2017; pp. 1–6. [CrossRef]
62. Mahbub, M. UAV Assisted 5G Het-Net: A Highly Supportive Technology for 5G NR Network Enhancement. *EAI Endorsed Trans. Internet Things* **2020**, *6*, e4. [CrossRef]
63. Khan, A.; Khan, S.; Fazal, A.S.; Zhang, Z.; Abuassba, A.O. Intelligent cluster routing scheme for flying ad hoc networks. *Sci. China Inf. Sci.* **2021**, *64*, 182305. [CrossRef]