Telecommunication Systems
https://doi.org/10.1007/s11235-022-00893-3

Quality enhancement in a mm-wave multi-hop, multi-tier heterogeneous 5G network architecture

Aftab Ahmed[1] · Muhammad Zakarya[1] · Xuan Liu[2] · Rahim Khan[1] · Ahmad Ali[3] · Ayaz Ali Khan[1,4]

Accepted: 28 February 2022

Abstract

Millimetre-wave ultra-dense high capacity networks are an important component of future 5G and 6G cellular systems since they are providing extremely high network capacity and speed to the end users. However, disparate types of users coexist in such scenarios, which can make the heterogeneous network unfair in terms of allocation of resources to various users based on their requirements. Therefore, a mechanism is required for effective spectrum sharing and to achieve overall system fairness. In this paper, an analytical model is suggested, which is based on a two-dimensional Markov state-transition diagram, to help set the parameter values to control the issuance of resources in coexistence layouts. A restriction approach is further implemented to gain a fair balance of the Grade-of-Service (GoS) for both user groups using the User Admission Control (UAC) mechanism. The developed mechanism restricts access to various channel resources for users with complete choice to give a greater probability of access to different users with limited resource options. Various levels of restriction are investigated in order to offer a balanced low-blocking probability performance to both user groups in order to improve the overall network fairness. Also, the proposed approach could provide a precise level of Grade-of-Service guarantee for both the user groups if sufficient flexibility is available within the whole network. Our simulations results along with the analytical model demonstrate that approximately 30% to 45% enhancement, in terms of grade of service (GoS), could be achieved in high to medium loads by restricting some of the users' flexibility. From the analytical model, it is concluded that the blocking of shadowed users are significantly reduced from 7% to 4.5% at high traffic loads. Moreover, the obtained results and findings are verified using a number of case studies and large-scale simulations.

Keywords Heterogeneous networks · Resource management · Ultra-small-cell · User diversity · Grade-of-service · Restriction

1 Introduction

A Cellular network or mobile network is defined as a radio network of various devices (preferably mobile) distributed over land areas called cells where every device is served by at least one fixed location transceiver or boaster or base station. In these networks, every cell is bounded to use a different set of frequencies from the neighboring cells which is primarily used to avoid interference among devices belongs to these cells. Additionally, these different frequencies provide guaranteed bandwidth utilization within a particular cell in the operational cellular networks. In these networks, data traffic across the networks is increasing significantly, day by

✉ Muhammad Zakarya
mohd.zakarya@awkum.edu.pk

Aftab Ahmed
aftab.ahmed.khan@awkum.edu.pk

Xuan Liu
yusuf@yzu.edu.cn

Rahim Khan
rahimkhan@awkum.edu.pk

Ahmad Ali
ahmadali@sjtu.edu.cn

Ayaz Ali Khan
ayazkhan@ulm.edu.pk

[1] Department of Computer Science, Abdul Wali Khan University, Mardan, Pakistan

[2] College of Information Engineering (College of Artificial Intelligence), Yangzhou University, Yangzhou, China

[3] Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

[4] Department of Computer Science, University of Lakki Marwat, Lakki Marwat, Khyber Pakhtunkhwa, Pakistan

Published online: 09 April 2022

day, which is due the increasing number of mobile users and demand for the mobile devices [1].

Data traffic demand in cellular networks is escalating at an exponential rate. In 2020, a typical mobile operator is projected to download approximately 1 terabyte of data, yearly [1]. The collective capacity demands from various sources such as, organic traffic growth, social networking, user-generated content, and machine-type associated sensors and devices (internet of things) will necessitate orders of scale capacity; and this may significantly increase, particularly, in future wireless networks. This data traffic evolution, from heterogeneous devices, also requires a paradigm change in designing the network architecture and resource or capacity provisioning [2]. The link capacity is approaching its theoretical limit and further growth in system spectral efficiency is only possible by increasing the node deployment density, as suggested in [3]. Mobile network operators are looking to facilitate and encourage subscribers to off-load their data from small-scale (macro) base stations to the substitute single/multi-hop small-cell networks and vice versa; which necessarily results in a basic heterogeneous network (HetNet). In general, a HetNet comprises different tiers of networks of multiple cell sizes/footprints and/or of various radio access technologies. A macro base station overlaying a multi-hop, ultra-small cell network is a simple illustration of a multi-tier HetNet [2]. From the perspective of a consumer and customer, each HetNet requires to offer ubiquitous network coverage, high throughput capacity, ultra-high data transfer rates, secure transmission of traffic data, always-switched-on and highly available network services, and always-connected in order to observe the best network user experience and service offerings [1,4]. This should be noted that some of these network services might be essential for certain latency sensitive network applications and quality of service (QoS).

While designing a HetNet system, we should also consider the probability that the network may not be connected, i.e. the probability of channel unavailability to consumers. However, various key factors such as different obstacles, degradation of signals, and user equipment restrictions will reduce the availability of some portion of the network to some of the users. In such situations, those limited users will face an inadequate network environment. A composite state with different levels of accessibility is expected to occur, when users having limited network choices (number of channels) and users with a more comprehensive set of connectivity options coexist in the same network environment. This is primarily because different users often have dissimilar terrestrial locations, elevation/azimuth angles and/or antenna equipment choices [5]. Meanwhile, when different users' types coexist in the matching coverage zone using resources from a shared resource pool, then the levels of utilization of resources for both users' groups will essentially affect each other. The presence of diverse users' types can even worsen the performance of the disparate users; which certainly make the management of the frequency spectrum a challenging task. Therefore, an approach is needed for user admission control in order to offer an appropriate Grade-of-Service (GoS), which is probability of a particular call that is either blocked or delay time is more than expected in a particular circuit network, for both users' types. The GoS metric is the probability of a data/voice packet being delayed longer than the pre-defined interval and is illustrated as a decimal fraction [6]. The Quality of service (QoS) metric is the measurement of the entire service performance, such as computer network, the performance observed by the network users. QoS is a quantitative metric for packet loss, throughput, transmission delay, etc.

In this paper, we propose a user admission scheme for heterogeneous networks with multiple radio access technologies. The goal is to ensure fairness and quality guarantees for users in two groups with different access options (privileged vs. shadowed) [7]. By restricting privileged users' access to the shared resources, fairness between the two user groups is achieved. Furthermore, we present a novel model for mmWave, multi-hop, multi-tier heterogeneous network. After an extensive presentation of the system, a mechanism to improve resource allocation fairness is proposed. The proposed mechanism is mostly evaluated based on the blocking probability experienced by "privileged" and "shadowed" users. Furthermore, this paper offers a novel solution (based on Markov chain) and an analytical approach to model scenarios, as described earlier. Finally, we have tested and evaluated the proposed model through case studies and large scale simulations in a couple of scenarios. Following are the major contributions of the work presented in this paper:

- an analytical model based on a two-dimensional state-transition diagram is developed, to help resource providers, in order to set the parameter values and control the issuance of resources in a multi-hop, multi-tier HetNet setup;
- a restriction mechanism is implemented to two different case studies in order to guarantee system fairness, in terms of quality of service, for each user group having different levels of network accessibility;
- an analytical model is suggested for the restriction mechanism given a couple of case studies; and
- the proposed technique is validated through large scale simulation using certain plausible assumptions and parameters.

The rest of this paper is organized as follows. We offer an overview of the related work in Sect. 2. Section 3 describes the multi-hop, multi-tier HetNet architecture along with its key elements. Section 4 explains the coexistence user scenarios and the implemented restriction mechanism. This is followed by the two case studies for analytical analysis in

Sect. 5. The large scale simulation results are presented in Sect. 6. Finally, Sect. 7 concludes the paper along with several directions for future research and investigation.

2 Related work

A restriction mechanism is investigated in [5], where the antenna directionality is restricted to some users. In this work, the restriction technique is applied to mm-wave, multi-hop, multi-tier HetNet scenarios where the users mainly suffer from radio link outage, caused by earthly obstacles, significant signal degradation or instead they represent users equipped with a simple fixed antenna system. From the perspective of a consumer, HetNets need to provide ubiquitous coverage, secure and high data rates, high throughput capacity, always-on, and always-connected-to-best-network user experience. The design of a HetNet system should also consider the probability that the network cannot establish a particular connection, i.e. the probability that no resource is available to users. However, various key factors including different obstacles, signal degradation and user equipment restrictions will reduce the availability of some portion of the network to some of the users. In such situations, those limited users will face an inadequate network environment. The main contribution of this work is to develop an access control approach for heterogeneous networks that will help to balance the traffic loads between different tiers having disparate groups of users. It is useful to achieve equal and controlled resource allocation with overall system capacity maximization to users having different degrees of accessibility choices [8]. Most research works in the HetNets domain are primarily related to single-hop femtocells, picocells, and relaycells overlaid on the edges of a macrocell for cell range expansion—as discussed in the existing state-of-the-art literature [3,7,9,10]. here are also exists rich literature that have discussed other issues like quality enhancement, allocation, quality of service, grade of service, and user access privileges [4,6,11].

In this work, it is extended one step further to multi-hop, ultra-small cells which are spread throughout the macrocell coverage area as shown in Fig. 1. These multi-hop links consist of sevral single-hop access links and dual-hop backhaul links which give the flexibility to respond to sudden changes in the network as well as to minimize the energy consumption of the entire network. These short range multi-hop links enable data to be directed from the users towards different aggregation points when some portion of the network is unavailable at low occupancy levels [12]. A typical multi-tier HetNet can offer both: (a) Quality-of-Service enhancement for individual users on a link-by-link basis' and (b) Grade-of-Service (GoS) improvement for joint users as a group, by taking advantage of user diversity in the network [5,13]. A novel mm-wave, multi-hop, multi-tier heterogeneous architecture is proposed as an ultra-dense network solution for future 5G cellular networks as shown in Fig. 1. This HetNet is composed of a multi-hop, Ultra-Small Cell Network (USCN) with an overlay of a Macro Base Station (MBS). The MBS provides conventional single-hop access connectivity to users on the ground in coexistence with the USCN in the same coverage area [14,15].

In this paper, we address the circumstances and situations in which the coverage area of a macro base station (MBS) is overlaying with highly directive mmWave base stations (USCN), creating a heterogeneous network. In this situation, some UEs, called privileged, have the possibility to connect to both kind of BSs and the remaining, called shadowed, only to USCNs. The shadowed UEs may have access to a smaller amount of resource blocks with respect to the privileged ones and these blocks can be also accessed by the privileged ones, meaning that the shadowed users have to compete with the privileged users to get the resource blocks. Consequently, the network is clearly unbalanced in advantage of a group of user and the network underperforms because the resources are not optimally allocated. The author's aim is to create a fair network by blocking the possibility for the privileged users of accessing to the USCN resources. The scenario investigated will be a typical real world configuration for 5G, but the study of the interactions among the base stations seem not to be much explored (perhaps relatively unexplored) in the existing literature [9,14,15].

3 Multi-hop, multi-tier heterogeneous 5G network architecture

3.1 System architecture

The network architecture investigated in this study is a modified form of Manhattan grid like topology with ultra-dense deployment of small base stations [16–18]. The multi-hop network can be further classified into a low frequency access network and a mm-wave single/dual-hop backhaul network. Figure 2 presents an in-depth view of the proposed network architecure which is predominantly based on dual-hop short distance links. It visualizes the intermediate relay points between backhaul communicating nodes. The main objective is to deliver high throughput capacity density with the least latency in the service area with minimum cost. The key elements of this network architecture are as follows [19]:

- *Access Base Station (ABS)* This is an inexpensive entity which is responsible for providing the access to the end users. A large number of ABSs are mounted below rooftop height on street-lamps, traffic lights and so on. They form a dense ultra-small access network and are placed

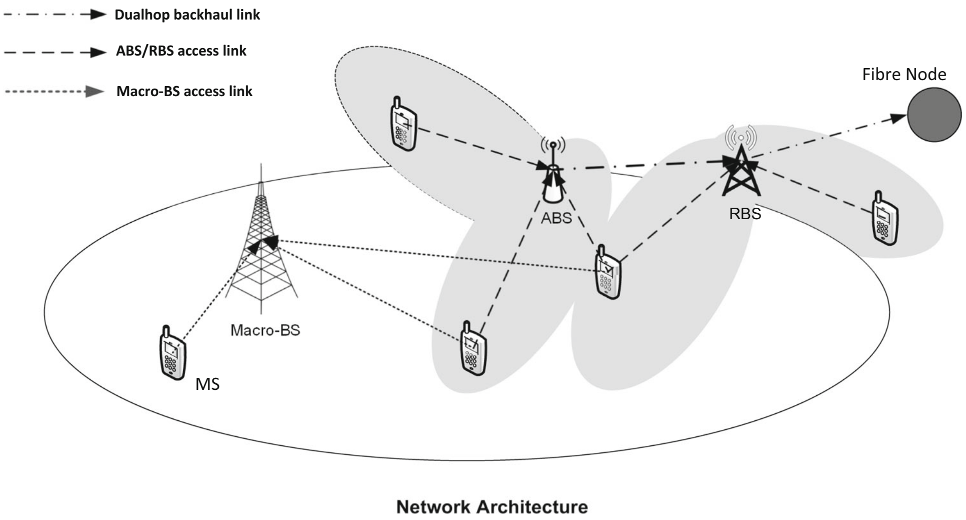


Fig. 1 Multi-hop multi-tier high capacity heterogeneous network [9]

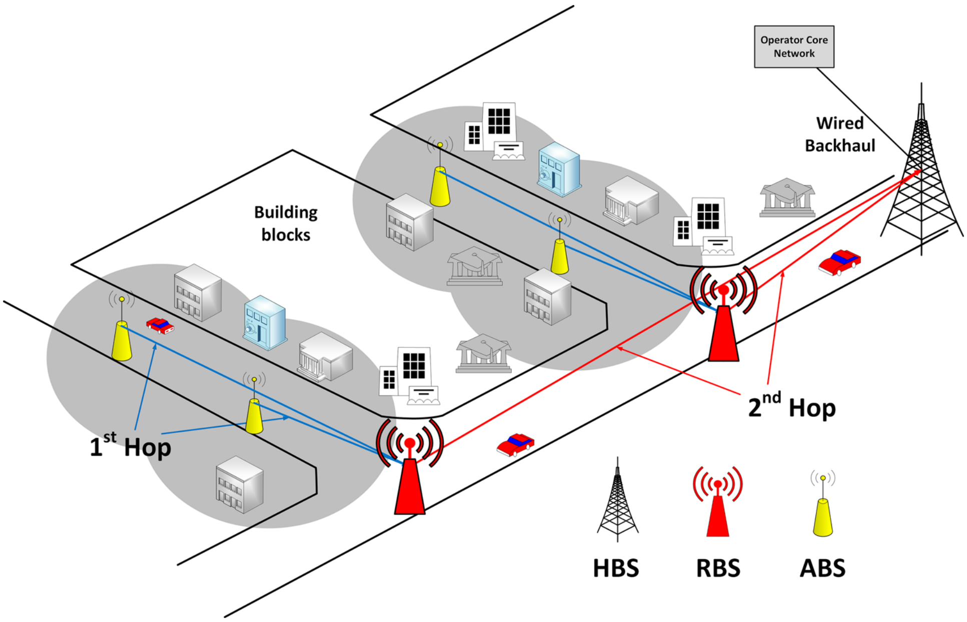


Fig. 2 A simplified dual-hop backhaul network model [9]

on street crossings. All these ABSs have four single wide-beam low frequency access antennas pointing along the streets to provide spatial diversity to outdoor MSs. These ABSs are also installed with narrow beam directional antennas for the dual-hop backhaul connectivity using millimetre waves. Furthermore, ABS are accessible directly to bother users to ensure single-hop or direct communication and each user is allocated a dedicated channel for communication.

- *Relay Base Station (RBS)* This is an intermediate node and is responsible for relaying the data traffic coming from the ABSs towards a Fibre Node. They are equipped with mmWave antennas and provide ultra-high capacity backhaul links for the uplink transmissions. Just like ABSs, these RBSs also have the same antenna design arrangements for its corresponding low frequency access network and are placed below rooftop heights on street lamps. Since, a direct access link to the RBS, preferably those reside in coverage area of a particular user, is provided as shown in the Fig. 1. This should be noted that we are not assigning a maximum number of RBs per user group, but we have proposed that if number of RBs are maximum then it is high likely that the performance of the system will be improved specifically in term of GoS [20]. Furthermore, the proposed scheme balances load of users via different available paths between ABS and FN.
- *Fibre Node (FN)* This entity is connected to the operator's core network via fibre and is placed on corners blocks. In this architecture, they act as an aggregation points where all the data from ABSs and RBSs are accumulated. It can be a bottleneck region for the ultra-small cell network because of data aggregation at this point. The ability to

serve a fixed number of distinct beams that are free from signal interference gives rise to this bottleneck.

3.2 Network path diversity

The graph $G(V, E)$, as shown in Fig. 3, represents the available path diversity for a single ABS node in the USCN. It illustrates the topology arrangement in the network for all the communicating nodes. $V(G)$ and $E(G)$ represent the set of vertices and directed edges respectively. A single ABS is in direct Line-of-Sight (LOS) of four RBSs and each RBS is then in LOS of two FNs placed one each on all 4 corners of a block.

As illustrated in Fig. 3, each ABS has a maximum of eight possible routes for transmitting its data to any aggregation point (FN) in the network. Thus, the existence of these multiple routes between all these nodes exploit the available path diversity in the network and result in much improved GoS and higher throughput capacity [16].

3.3 Dual-hop backhaul network

The multi-hop backhaul network links are mainly composed of single-hop and dual-hop connections. The data from/to multiple users is routed with low latency either via ABS/RBS to the Fibre Node through single/dual-hop mmWave backhaul links. A number of channel models have been used to calculate the path loss in LOS/NLOS scenarios, to effectively model real environments, including random effects such as attenuation due to shadowing [11,21]. Generally, wired networks or devices are difficult to handle then wirelessly connected devices particularly in scenarios where potential users are mobile entities. Note that a 2-hop set up is used in those scenarios where direct communication is not possible and it is primarily based on the distance measure, location information, and etc.

Both the backhaul links use the common mmWave resource pool due to its high directionality over a shorter distance. Also, the signal interference from one hop does not attenuate the signals on the second hop owing to enough spatial separation between them [12]. The Signal-to-Interference plus Noise Ratio (SINR) for RBS n on the 1st hop (signal transmitted from ABS m using resource block u) is:

$$\gamma_{n,u}^{m,l} = \frac{P_{B,m}^{A,l}\; g_u^{B,m,l,n}}{\sum_{i=1,i\neq m}^{N}\sum_{j=1}^{L} P_{B,i}^{A,j}\; g_u^{B,i,j,n} + \sum_{i=1,i\neq l}^{L} P_{B,m}^{A,i}\; g_u^{B,m,i,n} + \sigma^2} \tag{1}$$

where $g_u^{B,m,l,n}$ is the gain of wireless backhaul link from the l^{th} beam of the ABS m to RBS n. $\sum_{i=1,i\neq m}^{N}\sum_{j=1}^{L}\; P_{B,i}^{A,j}\; g_u^{B,i,j,n}$ is the signal interference from other ABSs to RBS n. $\sum_{i=1,i\neq l}^{L}\; P_{B,m}^{A,i}\; g_u^{B,m,i,n}$ is the interference from other beams of ABS m using the same resource block u. σ^2 is the noise power.

The link gain $g_u^{i,j}$ between two entities i, j is given by [22]:

$$g_u^{i,j} = \frac{G_i(\Theta_i)G_j(\vartheta_j)}{\overline{PL(d_{ij})}} \tag{2}$$

where $G_i(\Theta_i)$ is antenna gain of entity i along horizontal and vertical directions. $\overline{PL(\cdot p)}$ in Eq. 2 is the mean path loss that is assumed to have a linear dependence with the logarithmic distance, expressed as given by Eq. 3 [23]:

$$\begin{aligned}\overline{PL(d)}(dB) &= 20\log_{10}\left(\frac{4\pi d_o}{\lambda}\right)\\ &\quad +10\overline{n}\log_{10}\left(\frac{d}{d_o}\right) + X_\sigma\end{aligned} \tag{3}$$

where d_o is the close-in free space reference distance, $\overline{n}$ is the average path loss exponent. The carrier frequency is 75 GHz and X_σ is a Gaussian RV. The Signal-to-Interference plus Noise Ratio (SINR) from macro BS n on the 1st hop (signal transmitted from ABS m using resource block u) is:

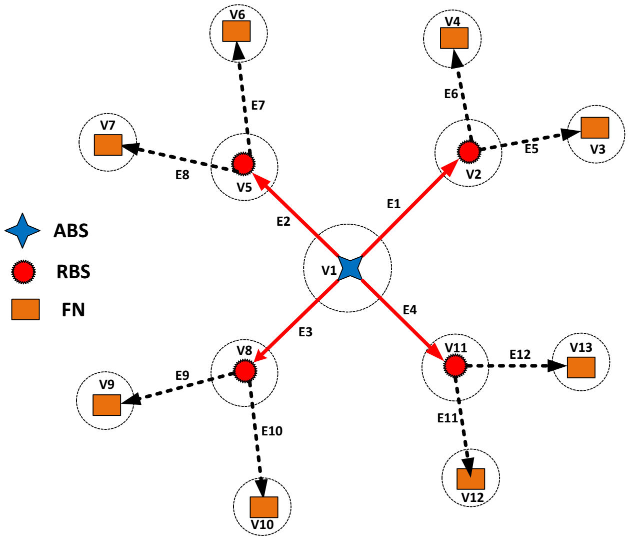


Fig. 3 Graph illustrating the topology for path diversity in USCN [12]

$$\gamma_{n,u}^{k,l} = \frac{P_{B,k}^{A,l}\ g_u^{B,k,l,n}}{\sum_{i=1,i\neq k}^{N}\sum_{j=1}^{L} P_{B,i}^{A,j}\ g_u^{B,i,j,n} + \sum_{i=1,i\neq l}^{L} P_{B,k}^{A,i}\ g_u^{B,k,i,n} + \sigma^2} \tag{4}$$

where $g_u^{B,m,l,n}$ is the gain of wireless backhaul link from the l^{th} beam of the macro BS k to ABS m and RBS n. $\sum_{i=1,i\neq m}^{N}\sum_{j=1}^{L}\ P_{B,i}^{A,j}\ g_u^{B,i,j,n}$ is the signal interference from other macro BS to ABS m and RBS n. $\sum_{i=1,i\neq l}^{L}\ P_{B,k}^{A,i}$ $g_u^{B,k,i,n}$ is the interference from other beams of macro BS k using the same resource block u. σ^2 is the noise power.

Similarly, the SINR metric for Fibre Node n on the second-hop (the signal transmitted from RBS m using resource block v) is derived in the same way as given in Eqs. 1–3. As, each Fibre Node covers two orthogonal streets, they are equipped with the two directional antennas, one for each street. The use of directional antennas on both the sides of backhaul link is a key factor as the interference between the beams of base stations are kept to a minimum level. The main reason behind the multi-hop approach is to increase the flexibility of the network in order to respond to the dynamic changes as well as to minimize the overall energy consumption of the network which is explained thoroughly in [12,24].

3.4 Single-hop access network

The Mobile Stations (MSs) are facilitated by this single-hop access network. In this study, only the outdoor MSs are considered, since in 5G it is expected that the MSs that are located indoors will be served by indoor infrastructure [25]. The SINR for ABS/RBS n (signal transmitted from MS k using channel w) is:

$$\gamma_{n,w}^{k} = \frac{P_{A,k}^{M}\ g_w^{A,k,n}}{\sum_{i=1,i\neq k}^{P}\ P_{A,i}^{M}\ g_w^{A,i,n} + \sigma^2} \tag{5}$$

where $P_{A,k}^{M}$ and $g_w^{A,k,n}$ is the MS transmit power and gain of access link respectively from MS k to ABS/RBS n. The factor $\sum_{i=1,i\neq k}^{P}\ P_{A,i}^{M}\ g_w^{A,i,n}$ is the signal interference from all other active MSs ($i \neq k$) in the network to ABS/RBS n using the same channel frequency w. For the access network, with the assumption of an omnidirectional antenna at the MS end, the link gain is obtained by:

$$g_w^j = \frac{G_j(\vartheta_j)}{PL(d_j)} \tag{6}$$

The channel propagation is modelled using the WINNER II B1 propagation model [26] for the low frequency (3.5 GHz) ultra-small cell access network as both the BSs and MSs are deployed outdoors. In the WINNER II models the propagation parameters may vary over time between the channel segments. Evaluation of small cell scenarios with outdoor and indoor users will require modification to the baseline model. PLlos of is the path loss of B1 LOS scenarios. The path loss for LOS scenarios is given by the following Eq. 7:

$$PL_{LOS} = \begin{cases} 22.7\log_{10}(d_1)+41+20\log_{10}\dfrac{f_c}{5} \\ 40\log_{10}(d_1)+9.45(d_{BP})+17.3\log_{10}(h'_{BS}) \\ -17.3\log_{10}(h'_{MS})+2.7\log_{10}\dfrac{f_c}{5} \end{cases} \tag{7}$$

where d_1 is the separation distance between MS and BS, f_c is the carrier frequency, d_{BP} is the breakpoint distance, h'_{BS} and h'_{MS} are the effective BS and MS antenna heights. The upper part of the Eq. 7 denotes the path loss for privileged users while the lower part represents the same metric for shadowed users. Meanwhile, the NLOS path loss (PL_{NLOS}) is given by:

$$PL_{NLOS} = min(PL(d_1, d_2), PL(d_2, d_1)) \tag{8}$$

where the $PL(d_1, d_2)$ can be computed using:

$$PL(d_k, d_l) = PL_{LOS}(d_k) + 20 - 12.5n_j + 10n_j\log_{10}d_l + 3\log_{10}(\frac{f_c}{5}) \tag{9}$$

while the value of n_j is the maximum value for all possible values of d_k; and is calculated through the following formula:

$$n_j = max(2.8 - 0.0024d_k, 1.84) \tag{10}$$

where $k, l\ \varepsilon\{1, 2\}$. Other important parameters mentioned in Eqs. 7–10 are further explained in [26]. During the uplink transmission, the effective signal strength at the receiver is obtained by accounting for the gains of MS and BS antennas, shadowing, path loss on the channel and interference from other users using the same resource blocks.

3.5 Macro-cell network

It is the second layer of this multi-tier heterogeneous network which is equipped with massive Multiple Input Multiple Output (MIMO) system. This MBS provides an overlay to all the

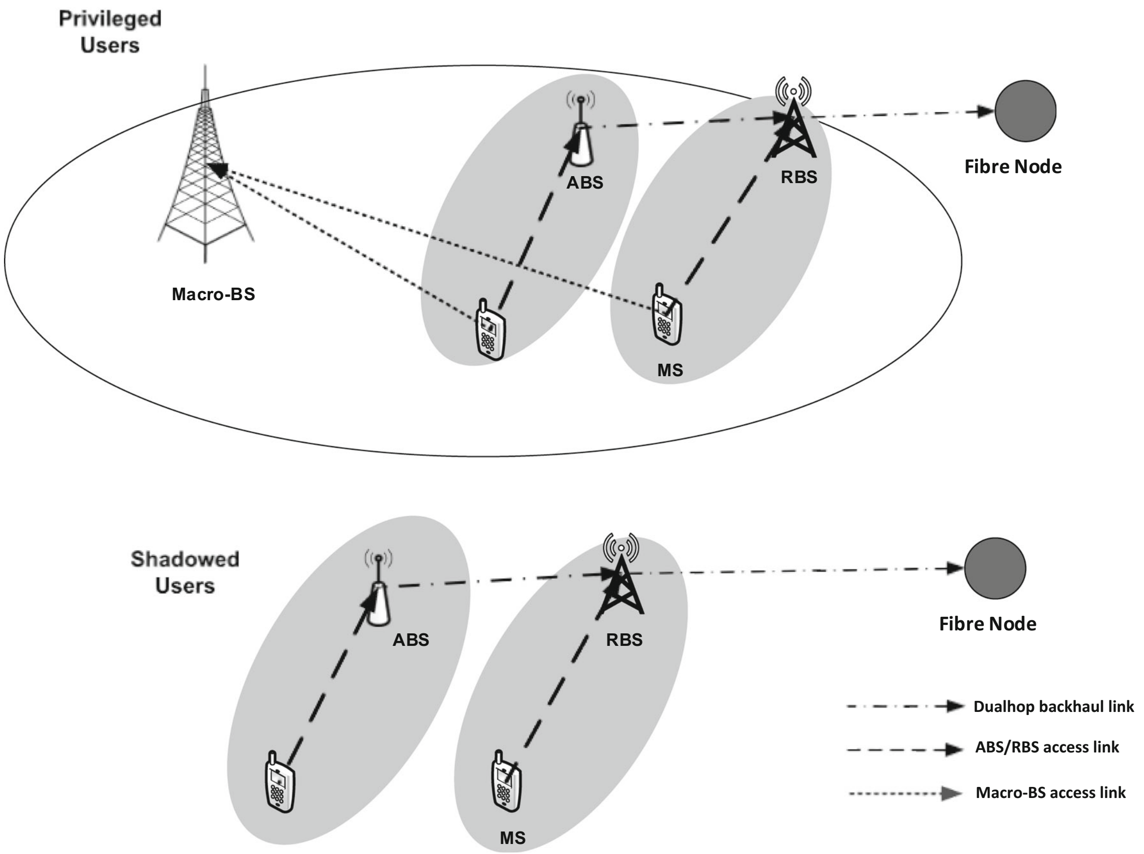


Fig. 4 A coexistence scenario with different types of user groups [9]

ultra-small cells in the same coverage area as shown in Fig. 1. Due to this system, it is possible to have an equal number of resource blocks for both the MBS access network and the ultra-small cell backhaul network. The radiation pattern of the 3-sector cell site is:

$$A(\theta) = -min\left[12\left(\frac{\theta}{\theta_{3dB}}\right)^2, \ A_m\right] \tag{11}$$

where $-180 \leq \theta \leq 180$. θ_{3dB} is the 3 dB beam-width and A_m is the maximum attenuation with further details given in [27]. The propagation model for this scenario is defined as:

$$L = 128.1 + 37.6 \log_{10}(R) \tag{12}$$

where R is the MBS-MS separation distance that is measured in kilometres (km). All other important parameters are listed in Table 3.

4 Coexistence scenario in multi-hop, multi-tier heterogeneous networks

A coexistence scenario is analysed for a different domain in [5], which dealt with users having restricted antenna directionality. In this work, we extend the applicability to different scenarios in a mm-wave, multi-hop, multi-tier HetNet domain where users suffer mainly due to an inadequate network environment. A range of elements including various obstacles, signal attenuation and MS terminal restrictions will lessen the accessibility of one tier of a network to some of its users. The restriction approach is developed further and analysed in a new domain for different types of users.

A diverse situation is probable to occur wherever users with a full access option and users with a restricted option coincide in the identical coverage area—as shown in Fig. 4. This is due to the fact that different MSs frequently have different locations (geographically distributed), elevation angles, and/or antenna equipment options [5,28]. The existence of various types of users can really reduce the performance of the overall system; and, thus, making the spectrum utilization process significantly inefficient. Generally, in cellular networks, users are divided into two groups (i) line of sight (ii) and those with obstacles. Now, obstacles can be of different type and may have different effect on the performance of the underlined network. However, for simplicity we have divided the potential users into two groups. Secondly, line sight users may be further divided into other group based on their distance, received signal strength indicator (RSSI), transceiver or transmitter capacity etc. However, if the valuable reviewer is insisting on to validate the proposed scheme in other scenarios as well then, we are happy to do so. The two user types which are investigated in this study are given below.

4.1 Privileged users

The privileged set of users can potentially access the MBS as well as the USCN as shown in Fig. 4. These have access to the resources available on both the tiers as shown in Fig. 5. In realistic scenarios, there are no obstacles to block this set of users from connecting to both the tiers due to their suitable geographical location and smart antenna systems. In addition, we are considering user mobility. We believe that user groups are not only based on their location information, but various other metrics are considered such as elevation angles and choice of the antenna equipment. Yes, user grouping is affected i.e., a direct line of site user may transfer to other group based on its current position and other metric. However, shifting of users from one group to another do not affect the performance of the proposed scheme.

4.2 Shadowed users

Shadowed users represent that set of users which are suffering from radio link outage, caused by earthly obstacles, significant signal degradation or instead they represent users equipped with a simple fixed antenna system. Therefore, this user group has limited resource options as shown in Fig. 5. Restriction of the privileged users up to some percentage is very useful as shown in the results section and it is due to the fact that these users have access to MBs and UCNS directly. Therefore, if restriction is not imposed then it will be difficulty (if not impossible) for a shadowed user to communicate. Furthermore, If privileged users are restricted completely then we believe that fairness will be compromised. We believe that the best possible scenarios for restrictions are 50 and 75% while other percentages achievers approximately less fairness that is why their results were not reported in the manuscript.

For GoS control in such scenarios, the MBS will require knowledge of load levels (i.e. the network traffic going through a BS) on single/dual-hop backhaul links which is primarily based on number of connected users and their communication activity preferably simultaneous. Every fibre node is bounded to have knowledge of information needed to computer load on a particular link, i.e., capacity of the channel, number of users (preferably simultaneous),l interference etc. This could be achieved through a Fibre Node sharing this information to the MBS through a dedicated control interface. Sharing of this information is very helpful in increasing GoS of the concerned networks as by knowing load levels probability of blocked and delayed user is decreased significantly. The Fibre Node and MBS then decide the access restrictions based on the detail whether the incoming user is a part of privileged or shadowed user groups. In this paper, through the restriction process, the GoS is controlled for both users' types [29].

4.3 Restriction mechanism

In order to remove the performance discrepancy, the restriction mechanism stops the privileged users from gaining access to the USCN. Meanwhile, the MBS is the only alternate option that they have to get a connection. In other words, we temporarily change the privileged users into a new type of restricted users by limiting their choice availability. Also, the shadowed users only have access to the USCN resources as shown in Fig. 5a. Note that the restriction approach, adopted in the paper, relies on an orthogonal partition of the available resources where a specific part of USCN resources is solely used by shadowed users. This partition was performed in the frequency domain. The main motivation, reason, and adoption of this kind of partition scheme with respect to the alternative underlay schemes is its simplicity. We are aware that the underlay schemes allow reuse of available resources, such as successive interference cancellation, advanced receiver designs, and etc.

In a controlled and flexible way, this approach prevents privileged users from accessing the USCN when their load levels is above a certain threshold limit as shown in Fig. 5b. The utmost elementary approach is the constant restriction function which put on an identical and same probability of restriction over the privileged users. It can be written as:

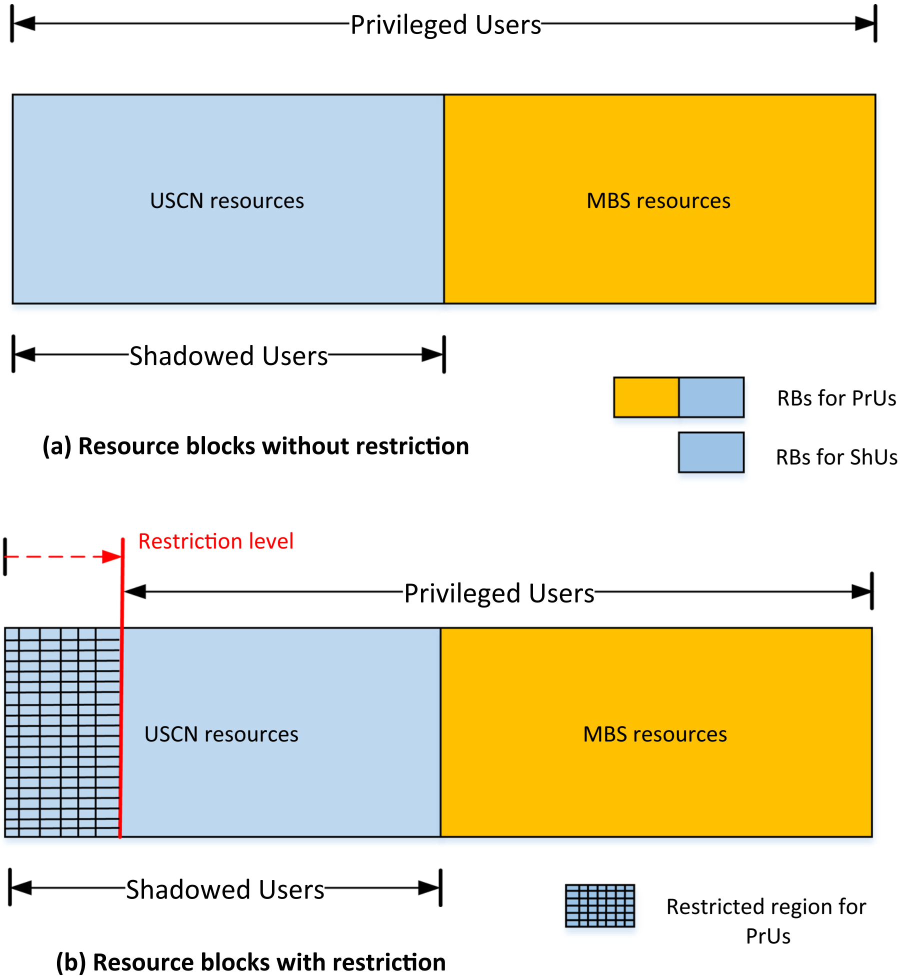
$$r(j) = C_c \qquad 0 < C_c < 1 \tag{13}$$

where C_c is the coefficient of restriction function. It reserves the reminder of the available resource blocks in the ultra-small cell network for shadowed users which have more limited choice flexibility in this wireless network [5]. However, once the load level of the USCN falls below the threshold set by Eq. 13, the temporary restriction which is applied earlier on privileged users is removed. It is this restriction mechanism that maximizes the system performance by limiting the choice availability for its users. Some more advanced restriction functions are discussed in [5] which are beyond the focus and scope of this research work.

5 Analytical model

In this section, network performance of the co-occurrence situation is investigated. In view of individual users get into the system, we assume arrival and departure as poisson processes

Fig. 5 Restriction approach utilized to pay-off the inferior performance i.e. GoS of shadowed users



for both Privileged and Shadowed users. Figure 6 depicts a state transition diagram (rate) to illustrate the performance of n-resource blocks in the two-tier heterogeneous network. We believe that the best possible scenarios for restrictions are 50% and 75% while other percentages achievers approximately less fairness that is why their results were not reported in the manuscript.

Individually, every node within the diagram represents a state. The foremost digit in the node represents the number of resource blocks employed on the MBS whereas the subsequent digit in the node stands for number of resources employed on the last hop of the USCN. This final hop is responsible for the bottleneck in the USCN due to data accumulation. Also, it is assumed for this analytical study that there is no retransmission in case of blocking and no signal interference. The arriving process for both types of users is constant and exponential distributed. Notations used in Fig. 6 are explained below in Table 1. For simulation simplification purposes, we consider the same number of resources (n) in each tier of the network and assume equal arrival rate on the MBS and USCN.

Table 1 List of mathematical notations and parameters used in the simulations

Parameters	Values
λ_P	Arrival rate of the Privileged Users to access the network.
λ_S	Arrival rate of the Shadowed Users to access the network.
n	Number of resource blocks on each tier of a network.
μ	The per channel departure rate is constant while the service time $(1/\mu)$ is distributed exponentially.

5.1 Key performance indicators

Ultra High Definition streaming media and cloud computing have started to gain in popularity. In addition to high data rates, they have generic requirements for low latencies due to the conversational nature. While these applications have a broad range of requirements in terms of capacity, latency and information loss, the common challenge is to support a large amount of mobile devices/sensors over a wide area without affecting the performance of other services. Based on the demands of the new applications and the ever growing number of mobile devices, industrial and research initiatives have identified a set of Key Performance Indicators (KPIs). Some of the KPIs most relevant to this thesis are:

- Latency in ms is the time it takes for a small data packet to be transmitted over the network from initial generation of data to its ultimate usable reception;
- Mobility in km/h is the maximum speed between a vehicle and a communicating node, at which the network is able to deliver the required QoS; and
- the blocking probability is the degree of the grade of service that a telecommunication system can provide. It is measured in percentage (%).

5.2 Case study # 1

This case study describes the best case scenario where both group of users are detached as much as possible. They only interact with each other once the MBS is fully occupied as shown by the bottom edge in Fig. 6. The resource allocation process is a birth-death process [30]. Note that, the transitions in a vertical direction denote both the arrival and departure process on the MBS, whereas the transitions in the horizontal directions denote both the arrival and departure process on the USCN. Furthermore, in the vertical direction, we assume that the arrival rate on the MBS is equal to λ_P. Similarly, in the horizontal direction, we assume that the arrival rate on the USCN is λ_S, when resources on the MBS are not fully occupied ($j_1 < n$). When resources on the MBS are fully utilized, i.e. $j_1 = n$, then the arrival rate in the horizontal direction equals $\lambda_P + \lambda_S$. This is due to the fact that when privileged users, initially arriving at the MBS, cannot provision any available resources on the MBS, only then will they access the USCN in search of free resource blocks. The total system arrival rate is split into equal halves for the two user groups (i.e. $\lambda_P + \lambda_S = \lambda_T$). Only μ is the per channel departure rate, as shown in Table 1, which might be potentially constant while the service time $(1/\mu)$ is distributed exponentially. However, the departure rate in any direction is equal to $k\mu$, where k denotes the total number of busy resource blocks of that state.

The restriction mechanism explained in Sect. 4.3 is modelled in such a way to improve the inferior GoS of shadowed users. In Fig. 6, it is illustrated that due to restriction, those states are restricted for the privileged users along the horizon-

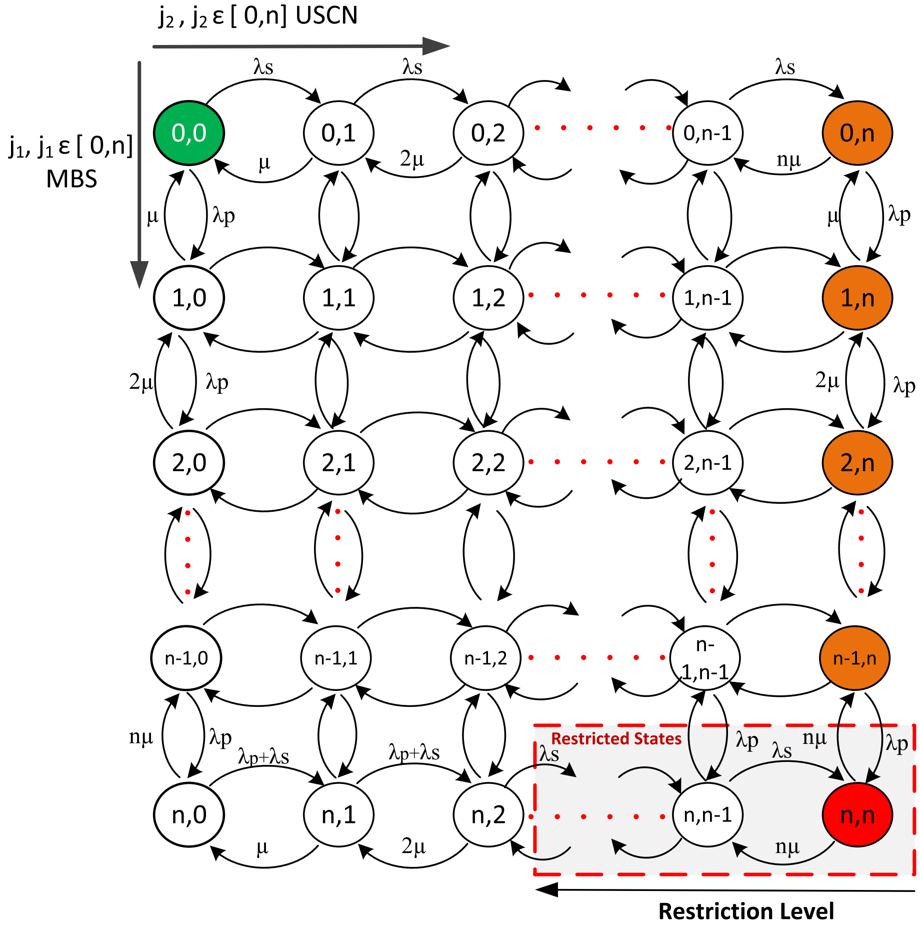


Fig. 6 The state-transition Markov model with the restriction approach

tal direction ($j_1 = n$). However, these states are accessible in a vertical direction by the privileged users. The restriction level extends from 0% all the way up to 100%. It is significant to govern and balance the distribution of resources to accomplish a reasonable provisioning pattern in the coexistence setup. In the Markov model we do not assume more than 1 RB to be allocated per user. It may be one or two or more depending on the deployment structure or position of RBs with respect to users. It is possible that a single RB may provide services to various users if other RBs are not deployed in their coverage area. The allocation of a potential RB depends on the required rate and channel of the requesting user in the cellular networks. Markov's analysis is a method for forecasting the value of a variable whose anticipated value is only impacted by its current state and no previous activity. In other words, it forecasts a random variable simply based on the existing circumstances around the variable.

The restriction process also equalizes the performance of both user groups. Using a flexible and controlled method, it blocks roughly privileged users to reserve appropriate network bandwidth for shadowed users that have restricted choice flexibility. It is this reimbursement effect which permits the network to attain a balanced blocking probability—as shown in Fig. 7. It is worth mentioning that the GoS performance of each user group is more important than any other criterion like the collective GoS of both the user groups. In this research study, the poor GoS of the shadowed users is due to the fact of their restricted access. The aim is to enhance the performance (GoS) of shadowed users at minimum expenses in terms of the entire system capacity, resource usage, and complexity.

5.3 Equilibrium analysis

Law of Conservation of Flow at statistical equilibrium states that the rate of flow into state (j_1, j_2) is equal to the rate of flow out of state (j_1, j_2). At any random point of time, the system can be in any state (j_1, j_2) with a state probability $P(j_1, j_2)$ [31]. The states in Fig. 6 could be divided into four different parts in the corners, six parts on the edge and only one part in the centre with total of eleven various equilibrium equation formats, respectively. However, it is impossible to have a non-positive number of provisioned resources in system; therefore, the condition $P(-1, j_2) = P(j_1, -1) = 0$, $(0 \leq j_1, j_2 \leq n)$ is applied to simplify the equilibrium expression. Using this method, the number of equation formats could, therefore, be minimized to only six, that are mathematically demonstrated as follows:

For state (j_1, j_2), $0 \leq j_1, j_2 < n$, we have

$$\begin{aligned}&(\lambda_P + \lambda_S + j_1\mu + j_2\mu) \cdot P(j_1, j_2)\\ &\quad = (j_1 + 1)\mu \cdot P(j_1 + 1, j_2) + (j_2 + 1)\mu \cdot P(j_1, j_2 + 1)\\ &\quad +\lambda_P \cdot P(j_1 - 1, j_2)\\ &\quad +\lambda_S \cdot P(j_1, j_2 - 1)\\ &[P(-1, j_2) = P(j_1, -1) = 0]\end{aligned} \tag{14}$$

For state (j_1, n), $0 \leq j_1 < n$,

$$\begin{aligned}&(\lambda_P + j_1\mu + n\mu) \cdot P(j_1, n) = (j_1 + 1)\mu \cdot P(j_1 + 1, n)\\ &\quad +\lambda_P \cdot P(j_1 - 1, n) + \lambda_S \cdot P(j_1, n - 1)\\ &[P(-1, n) = 0]\end{aligned} \tag{15}$$

For state (n, j_2), $0 \leq j_2 < R_L$,

$$\begin{aligned}&(\lambda_P + \lambda_S + n\mu + j_2\mu) \cdot P(n, j_2) = (j_2 + 1)\mu \cdot P(n, j_2 + 1)\\ &\quad +\lambda_P \cdot P(n - 1, j_2) + (\lambda_P + \lambda_S) \cdot P(n, j_2 - 1)\\ &[P(n, -1) = 0]\end{aligned} \tag{16}$$

For state (n, R_L),

$$\begin{aligned}&(\lambda_S + n\mu + R_L\mu) \cdot P(n, R_L) = (R_L + 1)\mu \cdot P(n, R_L + 1)\\ &\quad +\lambda_P \cdot P(n - 1, R_L) + (\lambda_P + \lambda_S) \cdot P(n, R_L - 1)\end{aligned} \tag{17}$$

For state (n, j_2), $R_L < j_2 < n$,

$$\begin{aligned}&(\lambda_S + n\mu + j_2\mu) \cdot P(n, j_2) = (j_2 + 1)\mu \cdot P(n, j_2 + 1)\\ &\quad +\lambda_P \cdot P(n - 1, j_2) + \lambda_S \cdot P(n, j_2 - 1)\end{aligned} \tag{18}$$

Finally, for state (n, n),

$$(n\mu + n\mu) \cdot P(n, n) = \lambda_P \cdot P(n - 1, n) + \lambda_S \cdot P(n, n - 1) \tag{19}$$

In the $(n + 1)^2$ equations as described above, only one is redundant, which justifies that it could be derived from other $(n + 1)^2 - 1$ equations. Since, the system will always be in a particular state; therefore, the state probabilities should essentially satisfy the normalization Eq. [30], given by the following Eq. 20:

$$\sum_{j_1=0}^{n} \sum_{j_2=0}^{n} P(j_1, j_2) = 1 \tag{20}$$

The $(n+1)^2$ equations along-with the normalization equation could be represented in a matrix format, given by:

$$AP = B \tag{21}$$

where P is the $(n + 1)^2$ x 1 state probability vector, A is the $(n + 1)^2$ x $(n + 1)^2$ coefficient matrix, and B is the $(n + 1)^2$ x 1 constant vector. Through solving the above equation of matrix, we can achieve the state probability vector i.e. P

and all other $(n+1)^2$ state probabilities i.e. $P(j_1, j_2), 0 \leq j_1, j_2 \leq n)$, effectively.

$$P = A^{-1}B \tag{22}$$

It is too complex to derive an expression of $P(j_1, j_2)$ in a closed form [32] for the restricted scenario presented in this paper. Therefore, we used numerically derived outcomes from the above equations in subsequent parts of the paper.

5.4 Blocking probability

The state probabilities i.e. $P(j_1, j_2), 0 \leq j_1, j_2 \leq n)$ that incorporate the restriction function are calculated. The blocking probability of Shadowed Users (PB_{SH}) is equal to the addition of the individual state probabilities on the right edge, as shown in Fig. 6.

$$PB_{SH} = \sum_{j_1=0}^{n} P(j_1, n) \tag{23}$$

Note that, the blocking probability of the Privileged Users (PB_{PR}) should also consider the blocking probability caused by the restriction mechanism along-with the state probability $P(n, n)$.

$$PB_{PR} = \sum_{j_2=R_L+1}^{n} P(n, j_2) \tag{24}$$

The restriction function also provides a certain degree of controllability over the whole network performance. For example, the probability of both the user groups can be equalized, i.e. For a complete fair network.

$$PB_{PR} = PB_{SH} \tag{25}$$

5.5 Analysis and results

In this section, the analytical model and the Monte-Carlo simulation results of the coexistence scenario for 24 resource blocks ($n = 24$) are investigated and compared. The arrival rates for both the user groups are identical ($\lambda_P = \lambda_S = \lambda_T/2$). Note that a 0% restriction means that the privileged users have complete access to the USCN as well as MBS while at 100% restriction level, the privileged users can only access the MBS. At 100% restriction, the whole USCN resources are reserved for the shadowed users.

As expected, with an increase of restriction level, the blocking probability for the shadowed users decreases while the blocking for the privileged users increases. From 60% restriction onwards, the improvement in performance is negligible due to system saturation. The restriction mechanism is able to equalize the performance of both the user groups as indicated by the fairness region in Fig. 7. These results also indicate the inherent discrimination in this system when both these users types coexist. In the case of no restriction, the shadowed users have much poorer performance compared to privileged users due to their confined number of choices.

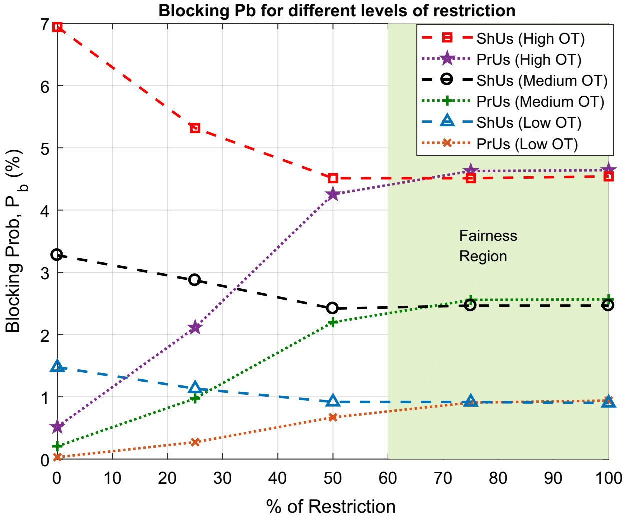


Fig. 7 Performance evaluation of restriction for certain levels of offered traffic [OT corresponds to λ_T]

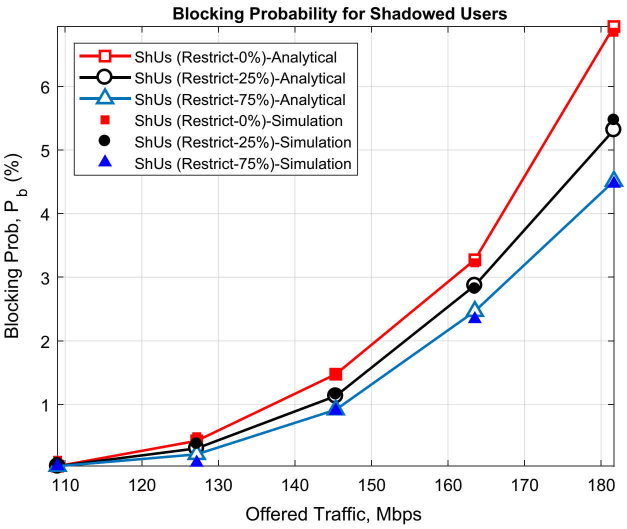


Fig. 8 Impact of increasing offered traffic on shadowed users

Figures 8 and 9 illustrates the effect of increasing offered traffic on the blocking probability performance of both the user groups. At 180 Mbps, the PB_{SH} is reduced from 6.9 to 5.4% and 4.5% with restriction levels of 25% and 75% respectively. From these results it is very clear that the restriction mechanism is mostly effective at high traffic loads. Therefore, it is most beneficial to postpone the restriction

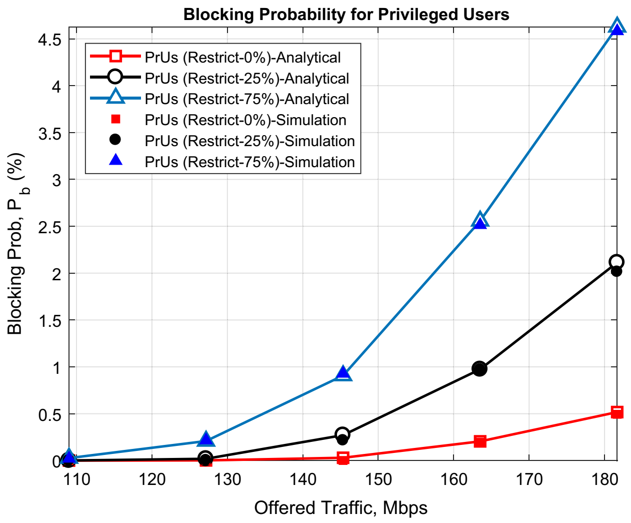


Fig. 9 Privileged users blocking for different levels of Offered Traffic

Table 2 List of mathematical notations

Parameters	Values
λ_{P1}	Arrival rate of privileged users to access MBS.
λ_{P2}	Arrival rate of privileged users to access USCN.

process until high offered traffic. The restriction is intended to provide free resource blocks for the shadowed users primarily in times of high traffic. For privileged users, the same restriction levels increases PB_{PR} from 0.4% to 2% and 4.5%. The slight discrepancy between the analytical and simulation results is due to the state probability $P(n, n)$ which represents the blocking caused by unavailability of resources rather than the restriction mechanism. Overall, these results demonstrate that the proposed analytical model is a good enough to represent the entire system.

5.6 Case study # 2

In the previous section, the case is analysed where the privileged users only access the USCN when there are no resources available on the MBS. However, this case study is more general and flexible where the performance of the restriction mechanism is analysed when the privileged users are allowed to have equal access to both types of networks from the very beginning. This approach is likely to be less advantageous due to greater interaction between both types of users. The aim is to understand the degradation in system performance which results from this higher degree of interaction. The set of equilibrium equations for this case study are detailed in Sect. 5.7. Two more notations for this case study are described in Table 2:

where λ denotes the arrival rate. The total arrival rate for such scenario is written as, $\lambda_T = \lambda_{P1} + \lambda_{P2} + \lambda_S$ while for privileged users, the two arrival rates ($\lambda_{P1} = \lambda_{P2}$) are equal to $\lambda_T/4$. Detail discussion is provided later in this section.

5.7 Equilibrium analysis

The set of equilibrium equations are given below for scenario (Case study # 2) when the privileged users are allowed to access the USCN from the very start. For state (j_1, j_2), $0 \leq j_1 < n, 0 \leq j_2 < R_L$, we have:

$$\begin{aligned}
&(\lambda_{P1} + \lambda_{P2} + \lambda_S + j_1\mu + j_2\mu) \cdot P(j_1, j_2) \\
&\quad = (j_1+1)\mu \cdot P(j_1+1, j_2) + (j_2+1)\mu \cdot P(j_1, j_2+1) \\
&\quad\quad + \lambda_{P1} \cdot P(j_1-1, j_2) + (\lambda_{P2}\lambda_S) \cdot P(j_1, j_2-1) \\
&[P(-1, j_2) = P(j_1, -1) = 0]
\end{aligned} \tag{26}$$

For state $(j_1, R_L), 0 \leq j_1 < n$,

$$\begin{aligned}
&(\lambda_P + \lambda_S + j_1\mu + R_L\mu) \cdot P(j_1, R_L) \\
&\quad = (j_1+1)\mu \cdot P(j_1+1, R_L) + (R_L+1)\mu \cdot P(j_1, R_L+1) \\
&\quad\quad + \lambda_P \cdot P(j_1-1, R_L) + (\lambda_{P2} + \lambda_S) \cdot P(j_1, R_L-1) \\
&[P(-1, R_L) = 0]
\end{aligned} \tag{27}$$

For state (j_1, j_2), $0 \leq j_1 < n, R_L < j_2 < n$,

$$\begin{aligned}
&(\lambda_P + \lambda_S + j_1\mu + j_2\mu) \cdot P(j_1, j_2) \\
&\quad = (j_1+1)\mu \cdot P(j_1+1, j_2) + (j_2+1)\mu \cdot P(j_1, j_2+1) \\
&\quad + \lambda_P \cdot P(j_1-1, j_2) + \lambda_S \cdot P(j_1, j_2-1) \\
&[P(-1, j_2) = P(j_1, -1) = 0]
\end{aligned} \tag{28}$$

For state (j_1, n), $0 \leq j_1 < n$,

$$\begin{aligned}
&(\lambda_P + j_1\mu + n\mu) \cdot P(j_1, n) = (j_1+1)\mu \cdot P(j_1+1, n) \\
&\quad\quad + \lambda_P \cdot P(j_1-1, n) + \lambda_S \cdot P(j_1, n-1) \\
&[P(-1, n) = 0]
\end{aligned} \tag{29}$$

For state (n, j_2), $0 \leq j_2 < R_L$,

$$\begin{aligned}
&(\lambda_P + \lambda_S + n\mu + j_2\mu) \cdot P(n, j_2) = (j_2+1)\mu \cdot P(n, j_2+1) \\
&\quad + \lambda_{P1} \cdot P(n-1, j_2) + (\lambda_P + \lambda_S) \cdot P(n, j_2-1) \\
&[P(n, -1) = 0]
\end{aligned} \tag{30}$$

For state (n, R_L),

$$\begin{aligned}
&(\lambda_S + n\mu + R_L\mu) \cdot P(n, R_L) = (R_L+1)\mu \cdot P(n, R_L+1) \\
&\quad + \lambda_P \cdot P(n-1, R_L) + (\lambda_P + \lambda_S) \cdot P(n, R_L-1)
\end{aligned} \tag{31}$$

For state (n, j_2), $R_L < j_2 < n$,

$$\begin{aligned}
&(\lambda_S + n\mu + j_2\mu) \cdot P(n, j_2) = (j_2+1)\mu \cdot P(n, j_2+1) \\
&\quad + \lambda_P \cdot P(n-1, j_2) + \lambda_S \cdot P(n, j_2-1)
\end{aligned} \tag{32}$$

Finally, for state (n, n),

$$(n\mu + n\mu) \cdot P(n, n) = \lambda_P \cdot P(n-1, n) + \lambda_S \cdot P(n, n-1) \tag{33}$$

5.8 Analysis and results

The total arrival rate for such scenario is written as, $\lambda_T = \lambda_{P1} + \lambda_{P2} + \lambda_S$ while for privileged users, the two arrival rates ($\lambda_{P1} = \lambda_{P2}$) are equal to $\lambda_T/4$. This restriction and assumption is applied in a similar way, as explained in Sect. 4.3, but the rate of transitions are different along both vertical and horizontal directions as indicated in Fig. 10. We are aware that different arrival rates will have an essential impact on our outcomes. Initially, when the load level on the USCN is below the restriction limit, the horizontal transition is the summation of λ_{P2} and λ_S but once that restriction limit is crossed, the horizontal transition is only caused by the λ_S as shown in Fig. 10. These are restricted states in the horizontal direction for the privileged users, but they are accessible vertically when those users want to access the MBS as indicated by vertical transition λ_{P1}. However, the remainder of the forward and backward transitions are the same as explained in Sect. 5.2.

Once, the resources on the MBS are fully occupied (i.e. $j_1 = n$), the states which are restricted by the restriction process become forbidden for the privileged users as shown by the bottom edge in Fig. 10. With such a configuration, there are not enough resources available for the shadowed users on the USCN due to severe competition between both the user groups and as a result the performance is severely degraded.

In this case, the resources available for the privileged users are significantly under utilized. Figure 11 clearly indicates that PB_{SH} is very poor even at low traffic loads but with application of the restriction mechanism, the performance of these users is enhanced. Without restriction, PB_{PR} is 0% due to the excessive availability of resources to privileged users. Also, above 60% restriction, the improvement in performance is negligible due to system saturation. From Figs. 11 and 12, it is clear that this kind of access for the privileged users makes the entire system underperform and inappropriate.

Figure 12 also illustrates that PB_{SH} is reduced from 9% to 4.2% with the restriction process. On the contrary, PB_{PR}

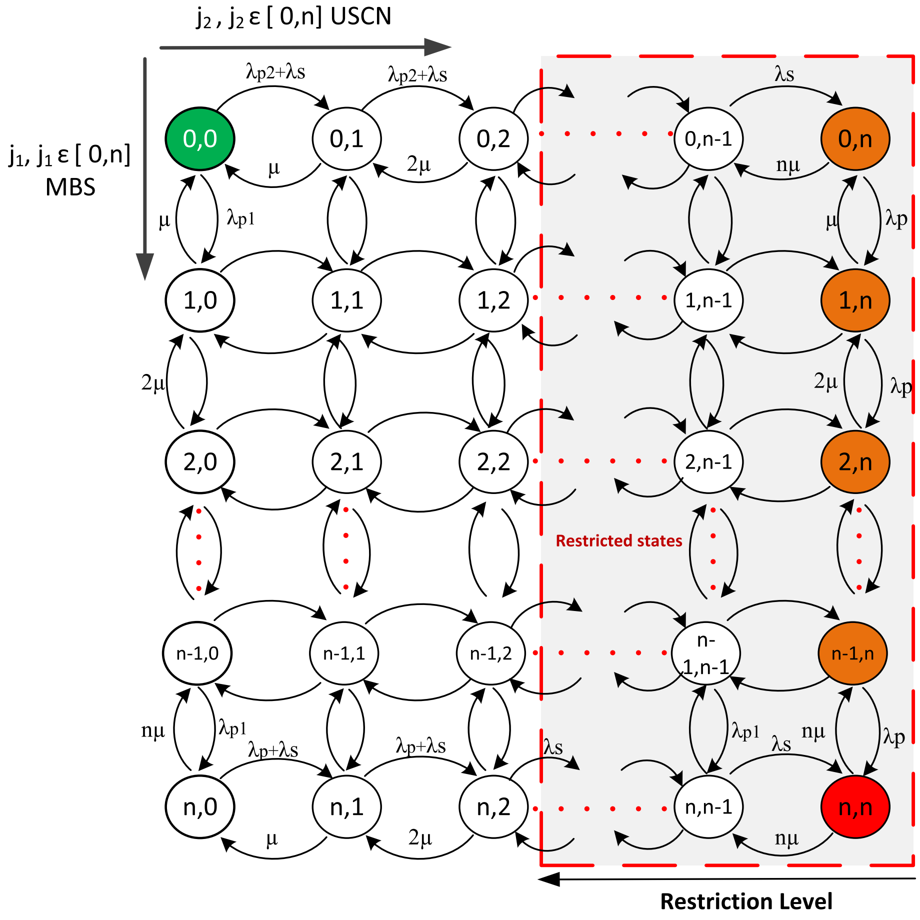


Fig. 10 The state-transition diagram in case of allowing a % of PrUs to access USCN from the beginning

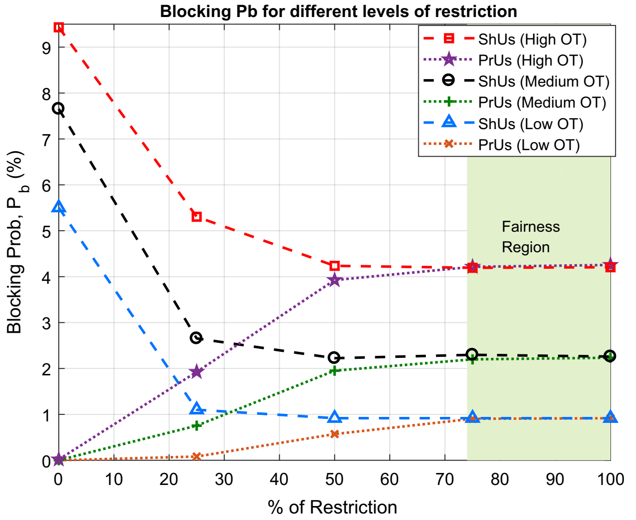


Fig. 11 Performance evaluation of restriction for certain levels of offered traffic [OT corresponds to λ_T]

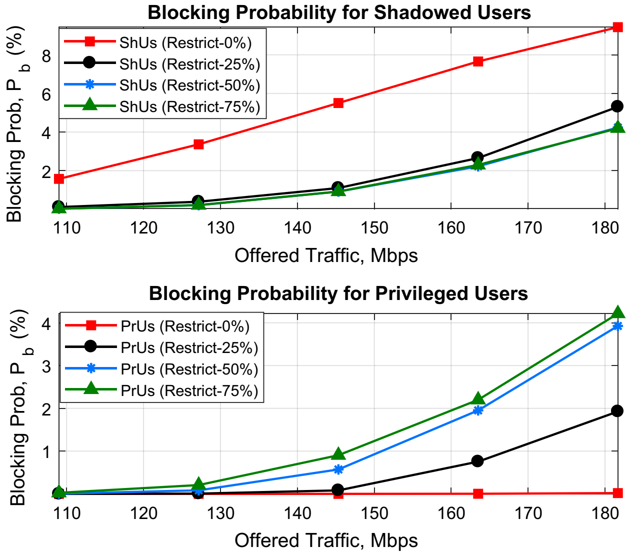


Fig. 12 Impact of increasing offered traffic on both users groups

is increased from 0% to 3.5% with 50% restriction. Again, it equalizes the blocking probability performance of both the user groups but the fairness region in this scenario is reduced. It is this compensation effect that allows the system to achieve a balanced GoS. Also, it is clear that the restriction process is most effective at high traffic loads. Thus, it is beneficial to suspend the restriction until high traffic levels because at low occupancy levels it is likely to cause unnecessary blocking. Therefore, the users with a high degree of choice are kept on those segments of a network which are inaccessible by the users with a more limited resource options. By comparing the blocking probabilities in the fairness regions of Figs. 7 and 11, for the different levels of OT, the convergence levels seem more or less the same, despite the different allocation scheme. This is due to the fact that both case studies have fewer variations, similar parameters, and most importantly, the outcomes are averaged over multiple runs. We observed significant differences between both for certain experiments. Furthermore, the blocking probability levels, when displayed in average, as shown in Fig. 13 are quite similar with those reported in Figs. 7 and 11. However, this closeness does not necessarily mean same outcomes [10,33].

6 Large scale simulation results

This section presents results which are obtained when a number of constraints, like signal interference and retransmission are included in the system. It is carried out to evaluate the performance of restriction mechanism explained in Sect. 4.3 on a large scale network with different levels of occupancy. The base station antenna profiles, gains and transmit powers are defined in [27,34]. Important simulation parameters are listed in Table 3. The experiments were performed in MATLAB2016 running on a server of 64 cores (3.8GHz) and 512GB of main memory. Various aspects of the network systems were modelled, using different parameters, as shown in Table 3, and described later in this section.

Table 3 Simulation and experimental parameters

Parameters	Values
Coverage area	450x450 m
Building size	75x75 m
Street width	15 m
Building height	6 m
Transmit power of RBS/ABS	35/35 dBm
Transmit power of MS for USCN/MBS	10/23 dBm
Antenna Height of FN/RBS/ABS/MS/MBS	10/4/4/1.5/15 m
Antenna Beams for Rx FN/RBS/ABS	2/3/4
Antenna Beams for Tx RBS/ABS	2/4
Antenna elements in MBS	8
SINR threshold for USCN/MBS	1.8/-3 dB
Log-normal shadowing factor	3 dB
Carrier frequency for USCN/BN/MBS	3.5/75/3.5 GHz
Number of FN/RBS/ABS/MS	4/16/16/2000
Resource blocks for USCN/BN/MBS	20/20/20
Inter-arrival time	Exponential Distribution
Noise floor	-114 dBm/MHz
Mean file size	2 MB
Iteration per offered traffic	200k

FN: Fibre Node, RBS/ABS/MBS: Relay/Access/Macro Base Station, USCN: Ultra-Small Cell Network, MBS: Macro Base Station, BN: Backhaul Network, MS: Mobile Station, Tx: Transmit, Rx: Receive

The main assumptions in this work are summarized below:

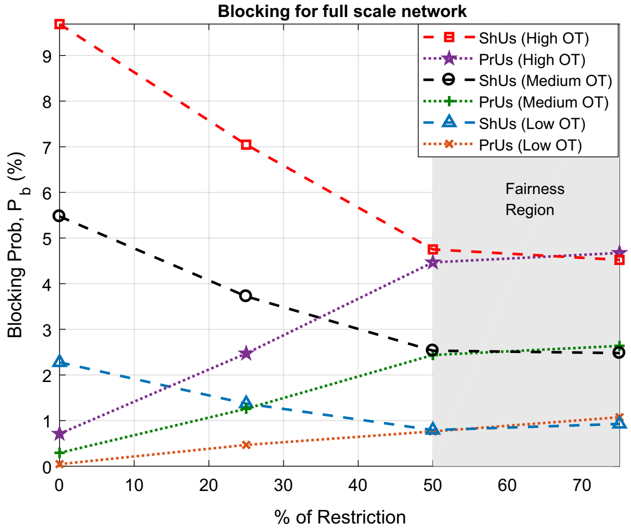


Fig. 13 Performance evaluation of restriction for all levels of offered traffic [OT corresponds to λ_T]

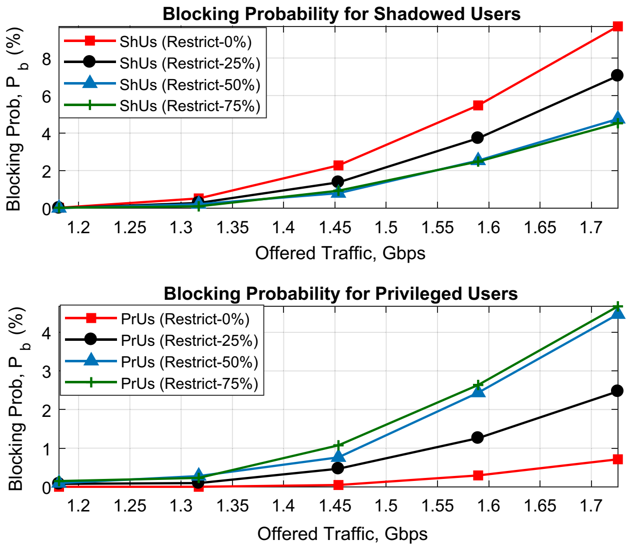


Fig. 14 Impact of increasing offered traffic on both users groups

- Data traffic is modelled using a file transfer-based traffic model, where the file size and the inter-arrival time follow an Exponential Distribution which simulates a succession of packets delivered in the network.
- It is assumed that the delay caused by queuing and signal propagation is negligible compared to the transmission time. For the uplink, specific mean to are employed to reduce the cubic metric of the transmitted signals; therefore, to improve the efficiency of transmitting power amplifier.
- Only the uplink transmission from the users are considered, as the data traffic likely to have a similar performance on the downlink side as well.
- Due to the small size of mm-wave antennas on the backhaul side of USCN, enough angular and spatial separation exists between the beams of ABSs/RBSs to minimize signal interference.
- Any blocked or interrupted files will be backed off for a random time and retransmitted until successfully delivered.

The metric used in this simulation to assess the full scale network performance is the blocking probability. It is worth mentioning that there are three main factors which give rise to blocking in the system: (i) unavailability of resources; (ii) the restriction process; and (iii) the poor quality of a communication channel.

Figure 13 illustrates the effect of increasing the restriction level on the performance of full scale network. In the absence of restriction, the blocking of shadowed users is severely high whilst for the privileged users, it is at very low level. As indicated in Fig. 13, the system achieves fairness around 50% of the restriction for all levels of occupancy. Furthermore, this is also clear that the restriction process is most beneficial at high traffic loads. Thus, it is appropriate to postpone the restriction until high occupancy because at low traffic loads it is likely to cause unnecessary blocking in the system. These simulation results also validate the analytical results mentioned in Sect. 5.5.

From Fig. 14, it is clear that in case of restriction, the rise in blocking for the privileged users is comparatively higher from the analytical model explained in Sect. 5.2. It is due to the presence of the signal interference and retransmission of traffic in the system. These simulation results also validate that the basic premise of applying the restriction factor is still applicable on a large scale scenario. Furthermore, it can be predicted from the analytical model what the restriction level needs to be for a complete fair system.

7 Conclusions and future work

In this paper, a novel mm-wave, multi-hop, multi-tier heterogeneous network is being investigated for the improvement in Grade-of-Service (GoS) and to achieve the overall system fairness and balance. The mm-wave, multi-hop, ultra-small cell and a single-hop massive MIMO base station collectively make the two-tier heterogeneous network. A restriction mechanism has been, then, applied to this multi-tier network which sacrifices some diversity of the mobile users for the sake of other user in the network. Furthermore, the users are divided into two separate users' groups. This classification of users is based on certain network parameters like geographical sites (location), azimuth/elevation angles and antenna equipment choices [35]. Different levels of restriction mechanism are examined and a comparison is drawn using certain performance metrics. From the analytical model, it is concluded that the blocking of shadowed users are significantly reduced from 7% to 4.5% at high traffic loads. It is achieved by limiting some choice flexibility of full access users for more restricted type of users. The restriction mechanism reserves some portion of ultra-small cell network for users with restricted choice flexibility, once the load level crosses that load threshold. It is this compensation effect that allows the system to achieve a balanced and fair GoS for the entire network [36,37].

From both the case studies, it is concluded that such a restriction technique will be useful in real systems, where users having greater freedom of choice can be directed on to other parts of the network, inaccessible by other users. It is also shown how the multi-hop, ultra-small cell network operates well in the presence of MIMO base station and provides the flexibility to adapt to the dynamic changes in future wireless networks. The work in this paper can be extended in three different ways. (i) Only two user groups have been investigated in this paper for examining the system fairness. One user group has association with the dense ultra-small cells network while the second user group has access to both the tiers of the network. However, we can analyse a third user as well which has only access to the MBS part of the network [38]. Initially, the restriction mechanism has been applied only in one dimensional, but we can apply the restriction along both the dimensions. (ii) Moreover, the data traffic generated by the privileged users has been divided equally between both tiers of HetNet. However, we can study the effect of splitting the same data traffic into any two unequal portions. (iii) Finally, a constant restriction function has been used to implement the restriction process for achieving network fairness and grade of service. However, different types of restriction functions can be studied that progressively constrains access to channels at high occupancy level [39,40].

In the future, we will investigate how other parameters affect the blocking probability, e.g., number of antenna beams, SINR threshold, different number of resource blocks, etc. For simulation simplification, we considered the same number of resources (n) in each tier of the network and assumed equal arrival rate on the MBS and USCN. However, heterogeneity should be considered as an important parameter. The way fairness metric is defined is somewhat abstract, in this work. A more rigorous definition needs to be explicitly provided, e.g., using the Jain's fairness index which has been widely used for fairness evaluation in resource allocation among various traffic flows. This should also be noted that the restriction approach, adopted in the paper, relies on an orthogonal partition of the available resources where a specific part of the USCN resources is solely used by shadowed users. This partition was performed in the frequency domain. We are aware that the alternative underlay schemes allow reuse of available resources, such as successive interference cancellation, advanced receiver designs, etc.; and should be considered as part of our future research.

Acknowledgements This research was financially supported by the Abdul Wali Khan University Mardan (AWKUM), Pakistan. The research was conducted during a doctorate level study (which was funded by the AWKUM's faculty development program) under the supervision of Prof. David Grace at the University of York, UK.

Author Contributions AA: Research, Methodology, Conceptualization, Writing - Original Draft, Software; MZ: Writing - Original Draft; Visualization, Data Curation, XL: Software, Writing - Review & Editing; RK: Visualization, Validation, Investigation; AA: Visualization, Writing - Review & Editing; AAK: Writing - Review & Editing, Revisions;

Data availability All authors have read the manuscript and agreed for the submission. Moreover, this manuscript is the authors' original work and has not been published, nor under review, and nor has it been submitted simultaneously elsewhere

Declarations

Conflict of interest All authors declare that they have no conflict of interest.

References

1. Agiwal, M., Roy, A., & Sexena, A. (2016). Next generation 5G wireless networks: A comprehensive survey. *IEEE Wireless Communications Surveys and Tutorials, 18*(3), 1617–1655.
2. Hu, Q. R., & Qian, Y. (2014). *Heterogeneous Cellular Networks* (2nd ed.). New York: Wiley.
3. Damnjanovic, A., Montojo, J., Wei, Y., Ji, T., Luo, T., Vajapeyam, M., et al. (2011). A survey on 3GPP heterogeneous networks. *IEEE Wireless Communications Magazines, 18*(3), 10–21.
4. Attiah, M. L., Isa, A. A. M., Zakaria, Z., Abdulhameed, M. K., Mohsen, M. K., Mowafakm, K., & Ali, I. (2019). A survey of mmWave user association mechanisms and spectrum sharing approaches: An overview, open issues and challenges, future research trends. *Wireless Networks, 26*(4), 2487–2514.
5. Liu, Y., Grace, D., & Mitchell, P. D. (2009). Exploiting platform diversity for GoS improvement for users with different high altitude platform availability. *IEEE Transactions on Wireless Communications, 8*(1), 196–203.
6. Hindia, M. N., Qamar, F., Ojukwu, H., Dimyati, K., Al-Samman, A. M., & Amiri, I. S. (2020). On platform to enable the cognitive radio over 5G networks. *Wireless Personal Communications, 113*(2), 1241–62.
7. Reddy, V.A., Stüber, G.L., Al-Dharrab, S., Mesbah, W., Muqaibel, A.H. (2021) Energy-efficient mm-wave backhauling via frame aggregation in wide area networks. *IEEE Transactions on Wireless Communications.*
8. Shafi, M., Molisch, A. F., Smith, P. J., Haustein, T., Zhu, P., De Silva, P., et al. (2017). 5G: A tutorial overview of standards, trials, challenges, deployment, and practice. *IEEE Journal on Selected Areas in Communications, 35*(6), 1201–1221.
9. Ahmed, A. (2018) *Enhancement in network architectures for future wireless systems*. PhD Thesis, University of York.
10. Li, R., & Patras, P. (2019). Max-min fair resource allocation in millimetre-wave backhauls. *IEEE Communications Magazine, 19*(8), 1879–95.

11. Aldubaikhy, K., Wu, W., Zhang, N., Cheng, N., & Shen, X. (2020). mmwave IEEE 802.11 ay for 5G fixed wireless access. *IEEE Wireless Communications, 27*(2), 88–95.
12. Ahmed, A., & Grace, D. (2016). Energy-aware topology management for 5G dual-hop ultra-high capacity backhaul networks exploiting path diversity. In *IEEE International conference on ubiquitous and future networks,* pp. 1020-1025.
13. Sun, Y., Peng, M., Zhou, Y., Huang, Y., & Mao, S. (2019). Application of machine learning in wireless networks: Key techniques and open issues. *IEEE Communications Surveys & Tutorials, 21*(4), 3072–3108.
14. Kalsotra, S., Kumar, A., Joshi, H. D., Singh, A. K., Dev, K., & Magarini, M. (2019) Impact of pulse shaping design on oob emission and error probability of GFDM. *2019 IEEE 2nd 5G World Forum (5GWF)*, pp. 226–231.
15. Togou, M. A., Bi, T., Dev, K., McDonnell, K., Milenovic, A., Tewari, H., & Muntean, Gabriel-Miro. (2020). DBNS: A distributed blockchain-enabled network slicing framework for 5G networks. *IEEE Communications Magazine, 58*(11), 90–96.
16. Ahmed., A. & Grace, D. (2015). A dual-hop backhaul network architecture for 5G ultra-small cells using millimetre wave. In *IEEE international conference on ubiquitous wireless broadband,* pp. 51-57.
17. Abolhasan, M., Abdollahi, M., Ni, W., Jamalipour, A., Shariati, N., & Lipman, J. (2018). A routing framework for offloading traffic from cellular networks to SDN-based multi-hop device-to-device networks. *IEEE Transactions on Network and Service Management, 15*(4), 1516–1531.
18. Liu, Yuchen, Hu, Qiang, & Blough, Douglas M. (2020). Joint link-level and network-level reconfiguration for urban mmWave wireless backhaul networks. *Computer Communications, 164,* 215–228.
19. Jiang, T., Zhao, Q., Grace, D., Burr, A. G., & Clarke, T. (2016). Single-state Q-learning for self-organised radio resource management in dual-hop 5G high capacity density networks. *Transactions on Emerging Telecommunications Technologies, 27*(12), 1628–1640.
20. Liu, Y., Chen, X., Niu, Y., Ai, B., Li, Y., & Jin, D. (2018). Mobility-aware transmission scheduling scheme for millimeter-wave cells. *IEEE Transactions on Wireless Communications, 17*(9), 5991–6004.
21. Saha, C., & Dhillon, H. S. (2019) On load balancing in millimeter wave HetNets with integrated access and backhaul. In *2019 IEEE global communications conference (GLOBECOM)* pp. 1-6.
22. Balanis, C. A. (1997). *Antenna Theory, Analysis and Design (2nd Edition)*. New York: Wiley.
23. Cassioli, D., Annoni, L.A., & Piersanti, S. (2013) Characterization of path loss and delay spread of 60-GHz UWB channels vs. frequency. In *IEEE international conference on communications (Wireless Communications Symposium)* pp. 5153-5157.
24. Saha, C., & Dhillon, H. S. (2019). Millimeter wave integrated access and Backhaul in 5G: Performance analysis and design insights. *IEEE Journal on Selected Areas in Communications, 37*(12), 2669–2684.
25. Andrews, J. G., Buzzi, S., Choi, W., Hanly, S., Lozano, A., Soong, C.K., & Charlie, Z. (2014). What Will 5G Be ?. *IEEE Journal on Selected Areas in Communications, 32*, (6).
26. Kyösti, P., Meinila, J., Hentila, L., Zhao, X., & Jamsa, T. (2007) WINNER II channel models v1.2, IST-WINNER D1.V1.2.
27. 3GPP TR 36.942 (2012) Evolved universal terrestrial radio access (EUTRA); Radio Frequency (RF) system scenarios, version 11.0.0 Release 11.
28. Katzis, K., Pearce, D.A., Grace, D. (2004) Fairness in channel allocation in a high altitude platform communication system exploiting cellular overlap. In *IEEE international conference on wireless personal mobile 8*, (1), pp. 106-111, Italy.
29. Ma, Z., Cao, J., Guo, Q., Li, X., & Ma, H. (2020) QoS-oriented joint optimization of concurrent scheduling and power control in millimeter wave mesh backhaul network. *Journal of Network and Computer Applications,* 102891.
30. Kleinrock, L. (1975) *Queueing Systems Volume 1: Theory*. John Wiley & Sons.
31. Blaunstein, N., Giladi, R., & Freedman, A. (2002). Unified approach of GOS optimization for fixed wireless access. *IEEE Transactions on Vehicular Technology, 51*(1), 200–208.
32. Cooper, R. B. (1981). *Introduction to Queueing Theory (2nd Edition)*. New York: North Holland.
33. Yang, G., Haenggi, M., & Xiao, M. (2018). Traffic allocation for low-latency multi-hop networks with buffers. *IEEE Transactions on Communications, 66*(9), 3999–4013.
34. Fisusi, A., Grace, D., & Mitchell, P. (2017). Energy saving in a 5G separation architecture under different power model assumptions. *Computer Communications, 105,* 89–104.
35. Kazi, B., & Wainer, G. A. (2019). Next generation wireless cellular networks: Ultra-dense multi-tier and multi-cell cooperation perspective. *Wireless Networks, 25*(4), 2041–2064.
36. Monemi, Mehdi, & Tabassum, H. (2020). Performance of UAV-assisted D2D networks in the finite block-length regime. *IEEE Transactions on Communications, 68*(11), 7270–7285.
37. Fan, B., Ramirez, D., Huang, L., Wang, Y., & Aazhang, B. (2018). A cross-tier scheduling scheme for multi-tier millimeter wave wireless networks. *IEEE Transactions on Wireless Communications, 17*(8), 5029–5044.
38. Dehos, C., Gonzalez, J., Domenico, A., Ktenas, D., & Dussopt, L. (2014). Millimeter-wave access and backhauling: The solution to the exponential data traffic increase in 5g mobile communications systems ? *IEEE Communications Magazine, 52*(9), 88–95.
39. Togou, M. A., Bi, T., Dev, K., McDonnell, K., Milenovic, A., Tewari, H., & Muntean, G.-M. (2020). A distributed blockchain-based broker for efficient resource provisioning in 5g networks. *2020 International wireless communications and mobile computing (IWCMC),* pp. 1485–1490.
40. Narsani, H. K., Raut, P., Dev, K., Singh, K., Li, C.-P. (2021). Interference limited network for factory automation with multiple packets transmissions. *2021 IEEE 18th annual consumer communications & networking conference (CCNC)*, pp. 1-6.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Dr. Aftab Ahmed received his PhD degree in Electronic Engineering from University of York, United Kingdom in 2019. He is a lecturer in Computer Science Department of Abdul Wali Khan University Mardan, Pakistan. His research work is related to improvement in performance in ultra-dense high capacity networks. His other research interests include radio resource management, topology management to improve system performance and overall energy efficiency in ultra-dense high capacity wireless networks and Machine Learning

Dr. Muhammad Zakarya received the PhD degree in Computer Science from the University of Surrey, Guildford, U.K. He secured the most competitive Australian scholarship i.e. IPRS, and University of Surrey studentship for his Ph.D. study. He has completed his postdoc experience, from March 2020 to March 2021, at the Khalifa University, UAE - a top ranked University. He is an Assistant Professor with the Department of Computer Science, Abdul Wali Khan University Mardan, Pakistan, since 2010. His research interests include cloud computing, mobile edge clouds, Internet of Things (IoT), performance, energy efficiency, algorithms, and resource management. He has deep understanding of the theoretical computer science and data analysis. Furthermore, he also owns deep understanding of various statistical techniques which are, largely, used in applied research. He is an Associate Editor for the IEEE Access Journal and has served as TPC member in various international conferences and workshops including CCGrid2021, GECON, ICCCI, UCC, and FIT. Furthermore, he is the Program Director of a leading research group i.e. iFuture and head of the CBI (Cloud computing, Big data, and Intelligence) Laboratory, at the Abdul Wali Khan University Mardan (AWKUM), Pakistan. Recently, AWKUM was ranked, as No. 1 University in Pakistan and 538th in the World, by the THE raking system. Dr. Zakarya has authored more than 40 research papers in CORE raked IEEE transactions, journals, and conferences. Dr. Zakarya is a Senior Member of the IEEE and was listed in the top 2% scientist list for the year 2020. He has produced several PhD students who are working in different universities and academia across the country.

Dr. Xuan Liu (S'11-M'17-SM'21) received Ph.D. degree in Computer Science and Engineering at Southeast University, China, and is currently a lecturer and master supervisor with College of Information Engineering (College of Artificial Intelligence) at Yangzhou University, China, and on-the-job postdoctoral researcher with School of Computer Science and Engineering, Southeast University, China. He is also serving as an Editorial Board Member of Computer Communications, an Associate Editor of TELS, IET Networks and IET Smart Cities. Furthermore, he served(s) as a TPC Member of ACM MobiCom, IEEE INFOCOM, IEEE ICC, IEEE Globecom, IEEE WCNC, IFIP/IEEE IM, IEEE NOMS, IEEE PIMRC, IEEE MSN, IEEE VTC, IEEE ICIN, IEEE GIIS, IEEE DASC, APNOMS, CollaborateCom, etc. Besides, he served as a Reviewer for 200+ reputable conferences/journal papers. His main research interests focus on ubiquitous communication and networking (Future Internet architecture, DAN, ICN, VLC, V2R).

Dr. Rahim Khan is currently a postdoc fellow at the UMS, Malaysia. Dr. Khan received his PhD degree in computer science from the Ghulam Ishaq Khan Institute (GIKI), Swabi, Pakistan. He is an Associate Professor with the Department of Computer Science, Abdul Wali Khan University Mardan, Pakistan. His research interest includes the Wireless Sensor Networks (WSNs) deployment, Internet of Thing (IoT), routing protocols, outliers' detection, techniques for congestion control, Decision Support System (DSS), vehicular ad-hoc networks, data analysis and similarity measures.

Dr. Ahmad Ali holds a PhD degree from the Department of Computer Science & Engineering at Shanghai Jiao Tong University, China. Before, that, he completed his MS from the same school. His current research interest deals with deep learning, big data analytics, data mining, urban computing, cloud computing, and fog computing. His research has been published in reputable Elsevier journals.

Dr. Ayaz Ali Khan is currently an Assistant Professor of Computer Science at University of Lakki Marwat, Pakistan. He holds a PhD degree in Computer Science from the Department of Computer Science, Abdul Wali Khan University Mardan, Pakistan. He completed his M.Phil (MS) in Computer Science from COMSATS Institute of Information Technology (CIIT), Islamabad, Pakistan. His area of research includes energy-aware and performance-efficient scheduling, resource allocation, placement and management, at datacenter level. Moreover, he has enough knowledge of distributed systems, optimisation, game theory and computer programming. His work has appeared in several prestigious IEEE transactions, and Elsevier journals of reputable venues.