IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 January 2024, accepted 20 January 2024, date of publication 23 January 2024, date of current version 30 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3357663

RESEARCH ARTICLE

Discontinuous Reception Based Energy-Efficient User Association for 5G Heterogeneous Networks

LOKESH SHARMA[1], JIA-MING LIANG[2], (Member, IEEE), AND SHIH-LIN WU[3,4,5], (Member, IEEE)

[1]UfiSpace Company Ltd., New Taipei City, Tucheng District 23674, Taiwan
[2]Department of Electrical Engineering, National University of Tainan, Tainan 701027, Taiwan
[3]Department of Computer Science and Information Engineering, Chang Gung University, Taoyuan 33302, Taiwan
[4]Department of Cardiology, Chang Gung Memorial Hospital, Taoyuan 33305, Taiwan
[5]Department of Electrical Engineering, Ming Chi University of Technology, New Taipei City 24301, Taiwan

Corresponding authors: Shih-Lin Wu (slwu@mail.cgu.edu.tw) and Jia-Ming Liang (jmliang@mail.nutn.edu.tw)

This work was supported by the National Science and Technology Council (NSTC) under Grant 111-2218-E-182A-001, Grant 109-2221-E-024-012-MY3, Grant NSTC 109-2221-E-024-013-MY3, and Grant NSTC 109-2221-E-182-038-MY3.

ABSTRACT The 5G Heterogeneous Network (HetNet) enables massive connections and diverse applications for User Equipments (UEs), leading to an exponential increase in data traffic. This surge results in extensive overloading at co-tier and inter-tier levels, inefficiencies in UE power usage at the cell edge, and complexities in Quality of Service (QoS) support. In this paper, we propose a scheme to address load balancing among Base Stations (BSs), while simultaneously improving power efficiency and ensuring QoS for UEs. The proposed heuristic scheme comprises three phases. The first phase optimizes the Discontinuous Reception (DRX) configuration parameters for UEs, the second phase evaluates overloading, and the third phase offloads data from overloaded BSs to other BSs. We categorize performance indicators into 1) User Performance Parameters (UPP), including DRX power saving, packet drop rate, and end-to-end delay, and 2) Network Performance Parameters (NPP), such as system throughput. To validate our scheme, we design an analytical model based on a two-dimensional continuous-time Markov chain (2D-CTMC) and a semi-Markov process. In addition, we implement a simulator to validate the scheme's performance. The results demonstrate that the proposed scheme significantly enhances performance in terms of UPP and NPP.

INDEX TERMS Continuous-time Markov chain (CTMC), discontinuous reception/transmission (DRX/DTX), heterogeneous network (HetNet), load balancing, quality of service (QoS), semi-Markov process.

I. INTRODUCTION

The fifth-generation (5G) delivers solutions, architecture, technologies, and standards to achieve wide mobile broadband, massive machine-type communication, and ultra-reliable communication with low latency in the coming decade [1]. According to [2], 5G radio access technology aims at providing peak data rates of 20 Gb/s, User Equipment (UEs) experienced a data rate of 100 Mbit/s (for wide area coverage cases e.g., in urban and suburban areas), 1 Gbit/s (in hotspot cases e.g., indoor areas), a spectrum efficiency improvement of 3×, support for up to 500 km/h mobility, 1 ms latency, a connection density of 10^6 devices/km^2, a network energy efficiency improvement of 100×, and an area traffic capacity of 10 Mb/s/m^2. In [3], the new enhancements of Carrier Aggregation/Dual Connectivity (CA/DC), the number of Component Carrier (CC)s has increased from 5 to 16 in the licensed and unlicensed spectrum, the maximum bandwidth goes up to 1 GHz. Provides early measurement reporting and a faster activation of secondary cells, which helps in reducing the latency. With this regard, 5G needs densification with different kinds of base stations (BSs) with disparate transmit powers and spectrum capabilities in dense-Heterogeneous Networks

The associate editor coordinating the review of this manuscript and approving it for publication was Chi-Yuan Chen.

TABLE 1. Acronyms.

Acronyms	Definitions
$5G$	Fifth generation
$2D-CTMC$	Two-dimensional continuous time Markov chain
$B\&D$	Birth and death
CA	Carrier Aggregation
CC	Component carrier
DRX	Discontinuous Reception
DC	Dual Connectivity
$DCPF$	DRX complex performance factor
DPC	DRX power consumption
$HetNet$	Heterogeneous Network
LB	Load Balancing
MBS	Macro base Station
MCS	Modulation and coding scheme
$MLWDF$	Modified Largest Weighted Delay First
NPP	Network Performance Parameters
PF	Proportional fairness
$PDCCH$	Physical Downlink Control Channel
QoS	Quality of Service
$RARF$	Relative Allocated Resources Factor
RB	Resource Block
SC	Small Cell
SBS	Small Base Station
SR	Serving ratio
$SINR$	Signal to interference plus noise ratio
UE	User Equipment
UA	User Association
UPP	User Performance parameters

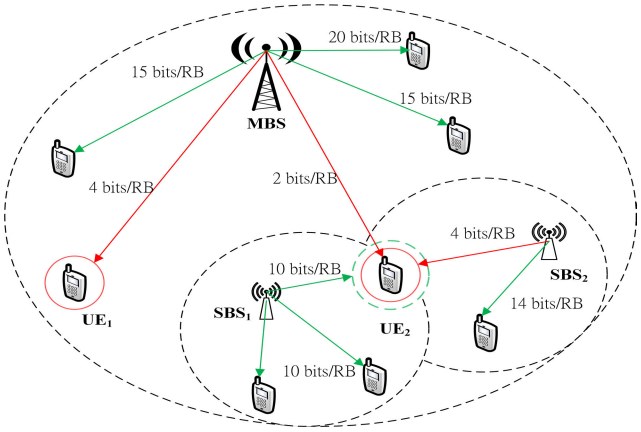

FIGURE 1. HetNet with poor UE performance.

(HetNet). Coexistence between 5G with other unlicensed technologies like Wi-Fi is challenging owing to the reduced probability of channel availability [4]. It creates an enormous problem of load balancing (LB) in the networks. LB can be solved by a combination of good user association (UA) and user-scheduling. The traditional UA, i.e., max-SINR (signal-to-interference-plus-noise ratio) makes UE prefer to connect to the macro base station (MBS) even though UE is within the coverage of small base stations (SBSs) which causes load imbalance in HetNet [5], [6], [7]. This causes the MBSs to be overloaded and the SBSs to be too underloaded. An UE connected to the overloaded high-power MBSs receives a poor data rate which needs more resource blocks (RBs) and higher power consumption to fulfill its application demand [8]. The overloaded MBSs do not satisfy the UEs' demands for *Quality of Service (QoS)* for different traffic flows in terms of traffic bit-rate, packet delay, and packet loss rate. To save the *power of UEs*, the defined *discontinuous reception (DRX)* mechanism allows UE to switch its radio interface on/off (fixed duration timers), based on data arrivals. The DRX coordinates with the BS and regulates UE to wake up periodically to receive/transmit data from/to the BS as shown in Figure 2. However, how to tune the DRX parameters to minimize energy i.e., *DRX optimization* is an open issue [9]. The current DRX exhibits power-saving issues at mmWave frequency [10]. The performance of DRX is degraded due to asynchronous DRX configurations, which is a major challenge in advanced technology such as DC [11]. Previously, our study [9] addresses the following challenges of guaranteed QoS and UEs' power consumption, and UA together. The asynchronous behavior is not addressed in this study. The definitions of the acronyms are provided in Table 1.

For example, Figure 1 shows the HetNet with one MBS and two SBSs, where the number of UEs associated to respective cells is based on max-SINR. UE_1 and UE_2 (Red color) receive poor performance in terms of bits per RB. To meet their requirements of QoS, the BS needs to allocate more RBs to UEs. However, more RBs require more power consumption of the UEs. On the other hand, UEs (green color) receive good performance. In all base stations covering the UE_2, SBS_2 is less congested than SBS_1, but the performance of RB is worse. For LB, UE_2 must be associated to SBS_2. For UE, the energy efficient association is SBS_1.

This research is driven by the need to address the challenges and shortcomings associated with User Association (UA). We have considered the UEs' QoS, power consumption, and load for UA. Our scheme coordinates between the UE power and the BSs load balance by offloading the UE with bad SINR from the MBS to other SBSs. While configuring the UE's DRX configuration parameters, we guarantee the QoS in traffic bit-rate, packet delay, and packet loss rate in the active traffic pattern. For each UE, we calculate the DRX complex performance factor (DCPF) and the Relative Allocated Resources factor (RARF), which gives higher priority to the UEs receiving less traffic bit-rate. Hence, a novel UA is proposed in the new HetNet environment, such that benefits both the network and the UE. In our study, for simplification, we categorize the performance indicators in two: User performance parameters (UPP) and Network performance parameters (NPP). The UPP is closely related to the traffic bit-rate, packet drop rate, end-to-end delay, and UE's power [9], [12], [13]. However, the NPP is related to the throughput [14].

Our research does UA while considering the UEs' power, and QoS. The proposed scheme coordinates between the UE power and the BSs load balance by offloading the UE with bad SINR from the MBS to other SBSs. Also, the optimization problem is also generalized to take into account the order in which many UEs are scheduled. In particular,

the main contributions of our work can be summarized as follows:

1) The scheme for DRX optimization problem, that configures the DRX configuration parameters of an UE, guarantee the QoS, such as traffic bit-rate, packet delay budget, and packet loss rate while decreasing the wake-up ratio of UEs.
2) The proposed scheme for UA, considers DCPF and the RARF, which gives higher priority to the UEs receiving less traffic bit-rate and giving benefits to both the network and the UE.
3) Proposed an analytical model based on a two-dimensional continuous-time Markov chain (2D-CTMC).

The rest of the paper is organized as follows. Section II, discusses the related work. Sections III and IV represent the System model for the HetNet environment and the proposed scheme. Section V discusses the Analytical performance model. Sections VI, and VII contain the Simulation results, and Conclusions, respectively.

II. RELATED WORK

Most earlier work on LB concerns to MBS only, under various joint situations such as BSs load, traffic load, channel state information, interference mitigation, instantaneous SINR [5], [14], [15], [16]. The QoS-based UA for better LB is discussed in [17] and [18]. Though some studies have considered a power for optimization such as [8] and [19]. However, the UE's power in joint situations is ignored while doing LB or UA. Therefore it is important to design a scheme that considers the UEs' power and QoS.

There are a lot of studies concerning the UA in HetNets under various joint situations. In study [15], authors have discussed the problem of static UA algorithms and re-association whenever a user arrived. The authors proposed an online algorithm motivated by online combinatorial auctions, which suggests the maximum number of potential association for UEs, that has provable performance guarantees. Simulations results claim that the proposed algorithms perform near-optimal and pose desirable fairness properties under realistic scenarios. Study [14] discussed the joint LB and interference mitigation in HetNet. The paper frameworks under the LB, the problem as network utility maximization subject to BSs load, traffic load, and channel state information. By stochastic optimization, the problem is decoupled into dynamic scheduling of MBS UEs, load provisioning of SBSs, and offloading MBS UEs to SBSs. The authors propose a hierarchical precoder to mitigate both co-tier and cross-tier interference, and achieve better performance in terms of SBS density, number of BS antennas, and transmit power levels, in cell-edge areas. However, the authors have investigated the results only with their proposed joint in-band scheduling schemes, without comparison to other scheduling schemes. Paper [16] investigates joint UA and user scheduling for LB and further improves the long-term throughput, motivated by the cumulative distribution function based scheduling. In which each BS implements its per-slot scheduling over its associated UEs, based on the instantaneous SINR, to select UEs for transmissions opportunistically, and exploit multi-user diversity. Study [5], proposed UA solution by distributed algorithm via Lagrangian dual decomposition achieves LB in HetNets, considering cell and RBs allocation jointly. A network utility maximization problem is formulated. The results have validated the performance of cell-edge UEs. However, in the above studies, authors have not considered the UE QoS parameters for LB, also they concluded that power control is an important technique to improve system performance, an interesting topic to investigate the joint power control, UA and scheduling for HetNets, which they will consider in their future work.

In these studies, some UAs, while considering the QoS, are investigated in HetNet. Study [18] designed the QoS-aware association by considering Proportional Fairness (PF) scheduling while minimizing end-to-end packet delay. The authors propose an optimal UA and resource allocation algorithm with the help of a classic knapsack problem. Simulation results minimize the average packet delay of UE traffic across the network. The goal of study [17] is joint optimization of RB and power allocated to UEs in HetNet. To maximize the number of UA and minimize the number of allocated RBs to meet the goal. Numerical results confirm that the developed scheme yields close-to-optimal performance. However, this study did not take into account the power saving of UEs, the delay of packet delivery, and the packet drop rate.

In the following studies, the authors consider power as an optimization parameter. Paper [8] has discussed the access network and backhaul energy consumption for the UA algorithm. The authors suggest energy-efficient heuristic UA by remaining RBs, the throughput of BS, and the data rate of associated UEs. The performance is evaluated in terms of energy efficiency and spectrum efficiency. Although, the results are not investigated in HetNet. Study [19] has discussed signal interference between BSs, proposing joint interference and power management. Based on the channel quality and traffic demand of each UE, it estimates the amount of RBs required by the UE. Further, management computes an almost blank subframes ratio, based on the principles of raising SBS GBR throughput and avoiding starving MBS UEs. To balance MBS and SBS loads and decides the DRX parameters accordingly. Simulation results show that the management achieves higher UEs' energy efficiency, higher network throughput, and lower packet dropping of real-time flows. According to traffic circumstances and threshold delays for 4G and beyond, the author of the study [20] suggests adapting sleep patterns. Results lead to significant power savings. However, in the overload scenario, no study shows the effects on the UPP has not been investigated.

In summary, to attain load balancing (LB) in the aforementioned user association (UA) schemes, the objective is

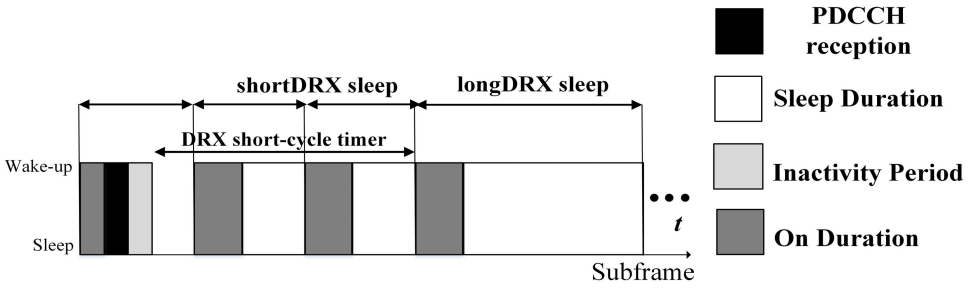

FIGURE 2. DRX mechanism.

to maximize either user performance parameters (UPP) or network performance parameters (NPP), but not both. Hence, motivated by the above limitations, we propose a novel scheme that can achieve efficiency, both UPP and NPP, to support the new HetNet.

III. SYSTEM MODEL

In this section, we discuss about the network architecture, determining the network capacity, and problem formulation. The notations that have been used in the system model are described in Table 2.

A. NETWORK ARCHITECTURE

In this study, we have focused on the UA and power saving model in the downlink of the 5G HetNet environment. The set of all macro and small BSs has been denoted by $\mathbb{B} \in \{1, 2, \ldots, B\}$ and dispersed in a HetNet network with respective transmit power levels; N available carrier frequency bands as a set of $\mathbb{N} \in \{1, 2, \ldots, N\}$ has been tuned to aggregate, each carrier $n \in \mathbb{N}$ and the set of distributed UEs has been denoted by $\mathbb{U} \in \{1, 2, \ldots, U\}$. Each BS has been tuned to the number of carrier frequencies. The CA/DC technologies allow an UE to associate with more than one number of BSs to increase the data rate. Currently, we assume that an UE is connected to two CCs of the same or different BSs and each CC corresponds to a serving cell.

Let $g_{i,j,n}$ be the power gain according to the Friis transmission equation as shown in Eq. (1), where $\ddot{T}(g_{i,j,n})$ and $\ddot{R}(g_{i,j,n})$ are the transmit and receive antenna gains from CC n, BS j to UE i, λ_n is the CC n wavelength, $d_{i,j}$ is the distance from BS j to UE i, d_0 is the far-field reference distance, and ξ is the path-loss exponent [21].

$$g_{i,j,n} = \frac{\ddot{T}(g_{i,j,n}) \times \ddot{R}(g_{i,j,n}) \times (\lambda_n)^2}{16\pi^2(\frac{d_{i,j}}{d_0}) \times \xi} \tag{1}$$

The SINR (η) of a CC to the UE is as shown in Eq. (2), where $p_{i,j,n}$ transmits the power from CC n, BS j to UE i, and σ^2 is the variance of additive white Gaussian noise [22]. The nominal bit-rate for an UE has been given according to Shannon's formula as shown in Eq. (3). Here W_n is the n^{th} CC bandwidth of a BS. UA balances the traffic load among the different BSs in HetNets and also optimizes the number of UEs. The fraction of RBs of BS j serves to UE i on CC n is $y_{i,j,n}$. Therefore, the overall long-term rate is shown in Eq. (4). Where $\sum_i y_{i,j,n} = 1$. Hence the nominal service rate for an UE is $\sum_j \sum_n \Phi_{i,j,n}$.

$$\eta_{i,j,n} = \frac{p_{i,j,n} \times g_{i,j,n}}{\sigma^2 + \sum_{k \in \mathbb{B}, k \neq j} p_{i,k,n} \times g_{i,k,n}} \tag{2}$$

$$\hat{\Phi}_{i,j,n} = W_n \times log_2(1 + \eta_{i,j,n}) \tag{3}$$

$$\Phi_{i,j,n} = y_{i,j,n} \times \hat{\Phi}_{i,j,n} \tag{4}$$

Some important terms that have been discussed here:
Requested RBs to a BS: It is the number of RBs, requested by the UE connected to a respective BS. The requested RBs of BS j, with CC n, with r requested RB by an UE i renders the equation as $\hat{R}_{j,n} = \sum_{i \in \mathbb{U}} r_{i,j,n}$. Hence the requested RBs in fraction to a BS for an UE are $\frac{r_{i,j,n}}{\sum_{i \in \mathbb{U}} r_{i,j,n}}$.

Allocated RBs by a BS: It is the number of scheduled RBs allocated by the BS to the UEs. The allocated load of BS j, with CC n, with y fraction of allocated RBs to the UE i by that BS shows the equation as follows: $R_{j,n} = \sum_{i \in \mathbb{U}} y_{i,j,n}$. Hence, the scheduled RBs associated with the function (such as PF) of the number of requested RBs by a BS for an UE show as $y_{i,j,n} = f(\frac{r_{i,j,n}}{\sum_{i \in \mathbb{U}} r_{i,j,n}})$.

Definition 1: An UE i is associated to CC n with the BS j, hence the overall long-term rate is $\sum_{i \in \mathbb{U}} y_{i,j,n} \times \hat{\Phi}_{i,j,n}$. Here $\sum_i y_{i,j,n} = 1 \ \forall j$, the total overall long-term rate for an UE i is the sum of different nominal service rates achieved by different CCs of BSs, as $\Phi_i = \sum_j \sum_n \Phi_{i,j,n}$.

The overall long-term rate, allocates proportional and equal RBs by giving fairness to each UE. Due to this, all the allocated RBs of a BS are divided proportionally as well as equally to the number of UEs associated with it. The logarithmic utility function achieves PF and trade-off between opportunism and unbiased allocation. It not only gives less priority to those UEs that already have high data rates for a long time range but also results in balancing the load. The optimization equation is as shown in Eq. (5). The indicator ($x_{i,j,n}$) corresponding to the association ($x_{i,j,n} = 1$, when UE i is associated with BS j of CC n, $x_{i,j,n} = 0$

TABLE 2. The notations used in the system.

Notation	Description
$\hat{\Phi}_{i,j,n}, \Phi_{i,j,n}$	nominal bit-rate and nominal service rate for UE i connected to BS j with CC n
$\hat{R}_{i,j,n}, R_{i,j,n}$	requested RBs and service RBs for UE i connected to BS j with CC n
$\mathbb{B}, B, j$	set of BSs, number of BSs, and indices
$\mathbb{N}, N, n$	set of CCs, number of CCs, and indices
$\mathbb{U}, U, i$	set of UEs, number of UEs, and indices
$g_{i,j,n}$	power gain
$\ddot{T}(g_{i,j,n}), \ddot{R}(g_{i,j,n})$	transmit and receive antenna gains
$d_{i,j}, d_0$	distance from BS j to UE i, and far-field reference distance
ξ	path-loss exponent
$p_{i,j,n}$	transmitted power
σ^2	variance of additive white Gaussian noise
W_n	bandwidth of a BS, with the CC n
$y_{i,j,n}, x_{i,j,n}$	fraction of RBs of BS, and association indicator
$Flow_i^{GBR}, Flow_i^{non-GBR}$	Guaranteed Bit Rate, and non-Guaranteed Bit Rate
f_j	application flow from the BS j
$\Phi_j^{GBR}, \Phi_i^{non-GBR}$	guaranteed bit-rate, and aggregate-maximum-bit-rate
$T_j^{delay}, T_j^{int-arr}, T_j^{rr}, T_p^{delay}$	packet delay budget, expected inter-arrival time, and service-request-response time
$Size_j^{min}, Size_j^{max}$	minimum packet size, and maximum packet size
χ, ϕ	total duration, and each DRX cycle duration
$T_i^S, T_i^L, T_i^{on}, T_i^{inact}, T_i^{st}, \kappa$	short, long, on-duration, inactivity timer, start offset, and start sub-frame of DRX cycle
Γ, Δ	DCPF, and RARF
$\tau_{j,n}, \Omega$	load value, and threshold load of each cell

otherwise).

$$
\begin{aligned}
&max_x \sum_{i\in\mathbb{U}} \sum_{j\in\mathbb{B}} x_{i,j,n} \times log(y_{i,j,n} \times \hat{\Phi}_{i,j,n}) \\
&s.t. \sum_{j\in\mathbb{B}} x_{i,j} = 1, \quad \forall i \in \mathbb{U} \\
&\quad x_{i,j} \in 0, 1; \quad \forall i \in \mathbb{U},\ \forall j \in \mathbb{B}
\end{aligned} \tag{5}
$$

B. DETERMINING THE NETWORK CAPACITY

In orthogonal frequency-division multiple access, the frequency domain is divided into 180kHz sub-channels and the time domain in slots of 1 ms. Such subdivisions are symbolized as scheduling RBs. Here each UE is connected to a number of applications or flows having some QoS parameters and the data packets from these flows arrive in the form of bursty traffic by the BSs. For better optimization, BS employs a scheduler that implements some scheduling schemes that utilize the number of RBs based on the used bandwidth, to the UEs. The allocation of RBs used by an UE depends on its SINR value. Due to this, the UEs' RB utilization also varies accordingly. The service rate (Φ) for each UE has been determined previously, whereas the network throughput ($\mathbb{T}$) for all the active UEs in the network is as followed by Eq. 6.

$$
\mathbb{T} = \sum_{i}^{U} \sum_{j}^{B} \sum_{n}^{N} \Phi_{i,j,n} \tag{6}
$$

C. PROBLEM FORMULATION

By utility function perspective, we assume that an UE i, while receiving nominal service rate $\Phi_{i,j,n}$ obtains utility $\psi(\Phi_{i,j,n})$, which is continuously differentiable, monotonically increasing, and strictly concave [23]. The aggregate utility function, which needs to be maximized. Since we assume that an UE can connect to more than one BS at a time, the indicator variable $x_{i,j}$ is of no use, whereas the $y_{i,j}$ remains valid. Hence, the objective is to allocate RBs to the number of UEs, with different rates to maximize the utility function (ψ) as shown in Eq. (7). It also provides an upper limit to the nominal bit-rate of the network. So, the joint association is as follows:

$$
\begin{aligned}
&max_y \sum_{i\in\mathbb{U}} \psi_i(\sum_{j} y_{i,j,n} \times \hat{\Phi}_{i,j,n}) \\
&s.t. \sum_{i\in\mathbb{U}} y_{i,j,n} \leq 1 \quad \forall j \in \mathbb{B} \\
&\quad 0 \leq y_{i,j,n} \leq 1, \quad \forall i \in \mathbb{U}, \forall j \in \mathbb{B}
\end{aligned} \tag{7}
$$

IV. THE PROPOSED SCHEME

To balance the load and enhance the performance of 5G HetNets, we propose a meta-heuristic scheme that optimizes the performance of UEs as well as the network. The scheme consists of three phases. The problem we address in the first phase is DRX optimization, providing a three-stage scheme, with the goal to determine each UE's DRX configuration parameters while guaranteeing the QoS parameters. In the second phase, we evaluate the network load balance scenario. Finally, in the third phase, we offload the UEs from the overloaded BSs to the underloaded BSs based on the results in the second phase. These scheme details have been described below:

A. FIRST PHASE: DRX CONFIGURATION

To calculate the DRX configuration parameters in downlink transmission, while a BS serves UEs, each UE_i acknowledges Guaranteed Bit Rate (GBR) $Flow_i^{GBR}$ and non-GBR $Flow_i^{non-GBR}$ flows. While each GBR flow f_j has guaranteed rate of Φ_j^{GBR} (bits/s), the non-GBR flow have aggregate-maximum-bit-rate $\Phi_i^{non-GBR}$ (bits/s). For each flow f_j, there is a packet delay budget of T_j^{delay} (ms) and the allowable packet loss rate of Φ_j^{loss}. The packet size ranges from $Size_j^{min}$ to $Size_j^{max}$ (bits/packet). Each flow f_j has an expected inter-arrival time of $T_j^{int-arr}$ (ms) and each non-GBR f_j has a service-request-response time of T_j^{rr} (ms) based on applications such as $T_j^{rr} \gg T_j^{delay}$. The UE_i channel rate $\Phi_i^{channel}$ varies over time. Let P_j be the number of packets of f_j that arrive during T_j^{delay}. Each packet $p, p = 1, \ldots, P_j$, poses a T_p^{delay} such as $T_p^{delay} = 1, \ldots, T_j^{delay}$. Let an UE remains in a connected mode for the total duration (χ), and each DRX cycle duration is ϕ. The shortDRX cycle (T_i^S) is the multiple of the smallest shortDRX sleep cycle, and the minimum of T_j^{delay} of UE, as shown in Eq. 8. Here T^S is the smallest duration of shortDRX sleep. The longDRX cycle (T_i^L) is the multiple of the shortDRX cycle, and the minimum of T_j^{rr} of UE, as shown in Eq. 9. The On-duration (T_i^{on}) is the smallest number of necessary wake-up durations in which all the high-priority packets can be received as shown in Eq. 10. The Inactivity timer (T_i^{inact}) is the maximum selected duration that satisfies the packet loss delay of all the flows, as shown in Eq. 11. It is based on UE's packet arrival within the T_j^{rr}, which has a higher probability, as shown in Eq. 12. Here κ is the subframe number that activates the T_i^{st}, and γ is the DRX start offset. The DRX Power Saving (DPS), DPC, and DAD have been estimated in Eq. 13, 14, and 15, respectively.

$$T_i^S = \left\lfloor \frac{min_j\{T_j^{delay}\}}{min\{T^S\}} \right\rfloor \times min\{T^S\} \tag{8}$$

$$T_i^L = \left\lfloor \frac{min_j\{T_j^{rr}\}}{min\{T_i^S\}} \right\rfloor \times min\{T_i^S\} \tag{9}$$

$$T_i^{on} = max\left\{ \left\lceil \frac{\sum_{T_j^{delay}=T_i^S, \forall f_j \in UE_i} Size_j^{max}}{\Phi_i^{min} \times RB} \right\rceil \right\} \tag{10}$$

$$T_i^{inact} = max\{T_P^{Delay} | P_j \in f_j\} \tag{11}$$

$$T_i^{st} = \left\lfloor \frac{max\{T_j^{rr}\}}{T_i^S} \right\rfloor \times T_i^S - ((\kappa - \gamma)\% T_i^S) \tag{12}$$

$$DPS_i = \frac{(T_i^S + T_i^L)}{\phi} \tag{13}$$

$$DPC_i = 1 - DPS \tag{14}$$

$$DAD_i = \frac{(T_i^S + T_i^L)}{\chi} \tag{15}$$

B. SECOND PHASE: EVALUATE OVERLOADING

In this phase, we have calculated two performance factors i.e. DCPF and RARF. The DCPF (Γ) is the product of DAD and DPC of an UE, that calculates the UPP, whereas the RARF (Δ) is the estimation for allocated RBs to the current and the neighboring CC's, which calculates the NPP for the network. To understand the above performance factors, we have first explained the single- and multiple-UE environment as follows:

Single-UE Environment: It considers a single UE in the network, which is the most favorable condition for the UE. Here an UE is allocated with the maximum RBs of a channel and defines the rate based on SINR. The DRX configuration parameters based on QoS get allocated to the UEs. Due to the random network deployment and inter-arrival packet durations, the average delay has a wide range and we have simplified them by normalizing, as shown in Eq. 16. Further, Eq. 17 shows the DCPF for the single-UE environment.

$$DAD_i^{norm} = \frac{DAD_i - \overline{DAD}}{max\{DAD\} - min\{DAD\}} \tag{16}$$

$$\Gamma_i^{single-UE} = DPC_i \times DAD_i^{norm} \tag{17}$$

Multi-UE Environment: It considers more number of UEs allowed to show competency for the RBs, which is the more realistic one. An UE is allocated RBs, according to the packet scheduling algorithm, which is deployed by the BS. Due to the realistic scenario, the ideal data rate gets degraded and other QoS parameters also get compromised. Equation 18 shows the DCPF for the multiple UE environment. To calculate the more accurate values of an UE's UPP, we have designed these environments.

$$\Gamma_i^{multi-UE} = DPC_i \times DAD_i^{norm} \tag{18}$$

DRX Complex Performance Factor (DCPF): It is the difference between the single-UE and multi-UE environment of DCPF, as shown in Eq. 19. A lower value suggests that the UE is enjoying favorable conditions, whereas the higher value shows that the UE is in unfavorable condition. It also shows the trade-off between delay and power consumption. Owing to this, we are able to select those UEs first that do not perform well in terms of UPP.

$$\Gamma_i = \Pi_i^{single-UE} - \Pi_i^{multi-UE} \tag{19}$$

Relative Allocated Resources Factor (RARF): It is the difference between the allocated RBs of the current and the neighboring BSs; which shows the impact of an allocated fraction of RBs to a BS, by the UEs. To offload prudently, we compare the current and neighboring cell's resource allocation. The UEs that show the significant change in the allocated RBs of the current BS reflected by the higher value of RARF, have been first.

$$\Delta_i = \left(\frac{y_{i,j,n}}{\sum_{i \in \mathbb{U}} y_{i,j,n}}\right) - \left(\frac{y_{i,j',n}}{\sum_{i \in \mathbb{U}, j' \neq j} y_{i,j',n}}\right) \tag{20}$$

These factors help us to select those UEs that perform severely so that the data of those UEs can be offloaded to other neighboring cells; which results in an increased NPP and UPP.

C. THIRD PHASE: TRIGGER THE OFFLOADING

The objective of this phase is to balance the load across the cells and improve UPP and NPP parameters. First, we set a threshold load value i.e. Ω for each cell, which is 75% in the paper [24]. The load value of each cell is ($\tau_{j,n}$) as shown in Eq. (21), which is the average service rate. If $\tau_{j,n} > \Omega$, it means a cell has already utilized its RBs and is currently in an overload condition. Similarly, the network load has been represented as τ. The fixed threshold value for the load usually shows adverse effects when the network is underloaded, and a cell is highly loaded. Therefore, to avoid such scenarios, we have designed a flexible threshold value estimation. To determine the overloaded cells, we consider the maximum value between the network and the cell loads, in Eq. (22). The load beyond this of any cell or network initiates our LB to balance the network load.

$$\tau_{j,n} = \frac{\sum_{i \in \mathbb{U}} R_{i,j,n}}{Total\ users\ of\ cell} \tag{21}$$

$$\Psi = max\{\tau, \tau_{j,n}\} \tag{22}$$

Finally, once the third phase gets triggered, those UE's (that perform badly for UPP and NPP) get selected from the overloaded cells in the network. In the second phase, the UEs that have a higher value of DCPF get identified and then offloaded to the underloaded cells based on RARF.

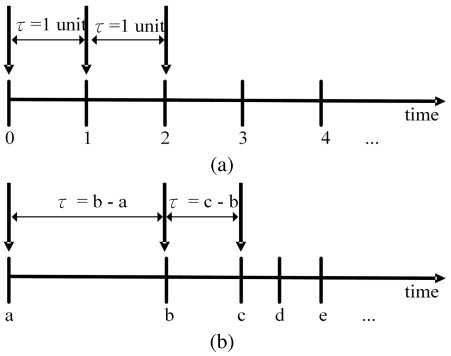

FIGURE 3. (a) Discrete Markov model (b) Continuous Markov model.

V. ANALYTICAL PERFORMANCE MODEL

A. NOTATIONS

We assume that a HetNet consists of $\bar{N}$ zones, each zone is disjoint, $1 < \bar{N} < \infty$. The zone is a representation of a cell in a BS. Here, $\bar{n} = 1, \bar{N}$ means that the variable $\bar{n}$, values in the set $1, \ldots, \bar{N}$, such that $\bar{n} \neq \bar{n}'$, $\bar{n}' = 1, \bar{N}$. A rate of transition of UE from the $\bar{n}th$ zone to $\bar{n}'th$ is represented as $\Phi_{\bar{n}\bar{n}'}$, when the rate of forced departure from the zone is represented as $\Phi_{\bar{n}-}$, and the rate of handover or departure to the adjoining zone is represented as $\Phi_{\bar{n}0}$. For simplification, we are not mentioning the specific UE, BS, and CC, while representing the Φ, $\hat{\Phi}$, etc. We only focus on the zones in this section.

B. GENERAL MODEL

In this section, we study a model to evaluate the performance of HetNet, based on a Continuous-Time Markov Chain (CTMC) [25]. The time Markov chain models are categorized in two: the Discrete-Time Markov Chain (DTMC) and CTMC, on the basis of fixed and variable time, respectively as shown in Figure 3. In DTMC, the event occurs at a known point of time, whereas in a CTMC, the event occurs at any point of time. The CTMC model is a stochastic process in which the state of the system can be viewed at any time, not just the discrete instants. Figure 4 shows the general analytical model for 2D-CTMC, in which Q represents the number of active UEs.

CTMC can be viewed as the Birth and Death (B&D) process, each arrival rate (λ) is represented as the birt.h, whereas the departure rate (μ) as death, and rates are exponentially distributed. Let $X(t)$ denote the state at time t, a CTMC process $X(t) : t \geq 0$ with state-space S. The state space S enters a state s, $s \in S$, continuous in time, then random variable H_s is the *holding time* of state s. The transition probability $P_{ss'}$, is the state transition from the state s to s', $s' \in S$. The holding time rate is the sum of the birth rates and death rates i.e., $\Theta_s = \lambda_s + \mu_s$. The formal definition is given by:

Definition 2: A stochastic process $\{X(t) : t \geq 0\}$ is called a CTMC, if for all $t \geq 0$, $t' \geq 0$, $s \in S$, $s' \in S$,

$$P(X(t+t') = s' | X(t) = s, \{X(u) : 0 \leq u < t\}) = P(X(t+t') = s' | X(t) = s) = P_{ss'}(t).$$

The transition probability $P_{ss'}(t)$ is from the current state s to the future state s'. The future, $X(t+t') : t' \geq 0$, given the present state $X(s)$, is independent of the past $X(u) : 0 \leq u < t$. Such a process is called a CTMC. The exponentially distributed holding time shows the memoryless property. The process enters in state s remains there independent of the past, for an amount of time $H_s \sim exp(\Theta_s)$. A CTMC can be described by a transition matrix $P = (P_{ss'})$, representing state transitions, with the holding time rates (Θ), with a set of rates $\Theta_s : s \in S$, completely determines the CTMC.

Definition 3: Whenever $X(t) = s \geq 1$, the next transition will be a birth with probability $P_{s,s+1} = P(B < D) = \frac{\lambda_s}{\lambda_s + \mu_s}$, and a death with probability $P_{s,s-1} = P(D < B) = \frac{\mu_s}{\lambda_s + \mu_s}$.

For each state $s \geq 0$, birth rate λ_s and death rate μ_s: Whenever $X(t) = s$, independent of the past. The time until the next random variable birth rate, and death rate which is $B \sim exp(\lambda_s)$, and $D \sim exp(\mu_s)$, respectively. Hence the holding time rates i.e., $\Theta_s = \lambda_s + \mu_s$, and holding time is given by $H_s = min\{B, D\} \sim exp(\lambda_s + \mu_s)$. When $\lambda_s = 0$,

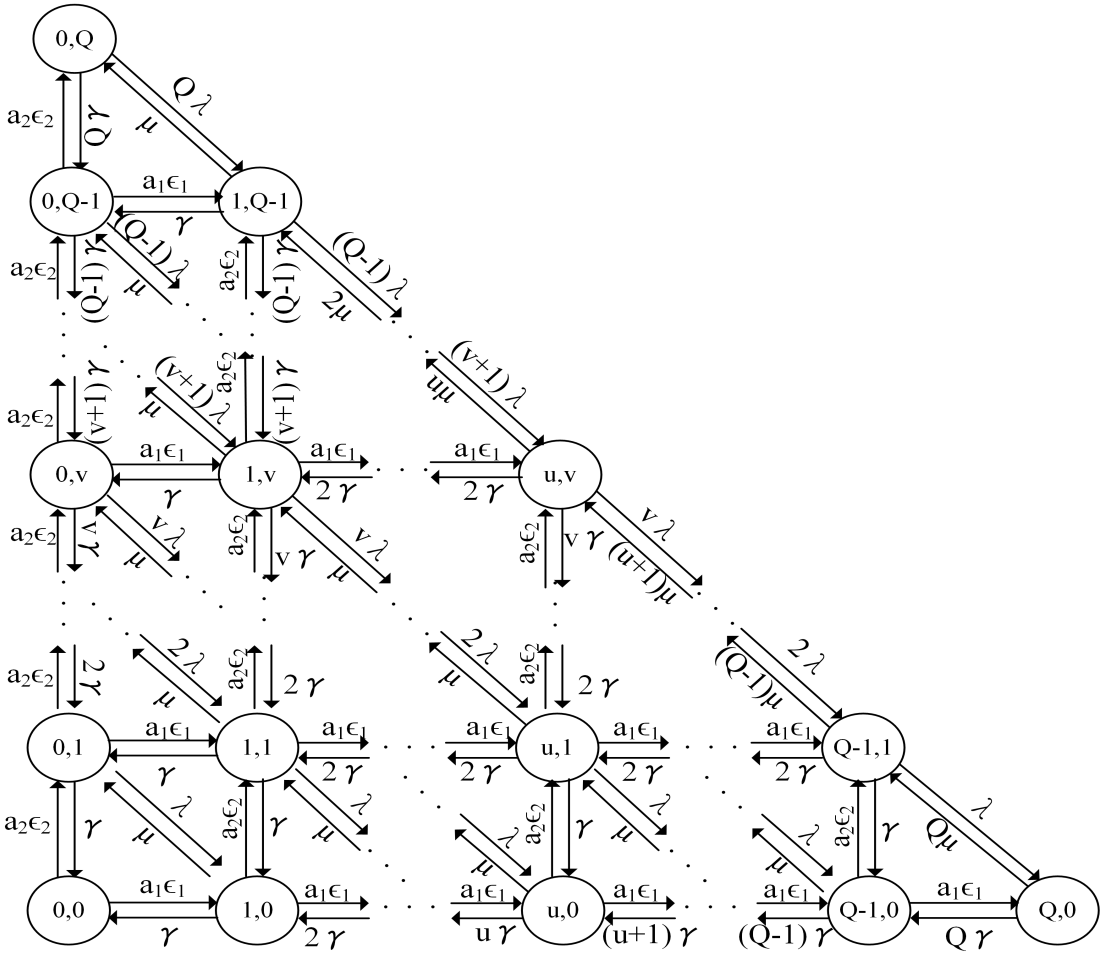

FIGURE 4. Two-dimensional CTMC model for Q active number of UEs [25].

$s \geq 0$, and $\mu_s = 0, s \geq 0$, we call the process is a pure birth process. The B&D process is a random walk, with dependent increment probabilities.

C. ANALYTICAL MODEL

1) MULTI-DIMENSIONAL CTMC

The HetNet consists of $\bar{N}$ zones. In general, zones are in concentric ring form, while in practice they can have various structures, due to the interference of nearby BS. In our study, an UE in zone $\bar{n}$ is not limited with zones $\bar{n}-1$ of $\bar{n}+1$ for adjustment. We propose a model using $\bar{N}$ dimensional CTMC, the state vector can be described as $\bar{N}$-tuple $S = (s_1, \ldots, s_{\bar{N}})$. Here S is the set of feasible states and s defines the number of UEs in zone $\bar{n}$, and the total number of zones is $\bar{N}$. The number of available scheduling RBs in zone $\bar{n}$, is represented by $\hat{R}_{\bar{n}}$. The term $f \in [0, 1]$ denotes the RBs that can be used for new UEs. If $f \times \hat{R}_{\bar{n}}$ is less than the threshold RB, the session $(a_{\bar{n}}(s))$ for the new user gets accepted; else rejected as shown in Eq. (23). As for how long an UE remains in a zone is a continuous-time interval (not the discrete one), which can be changed by an UE, based on QoS. Here an UE can have multiple choices to join other zones for better performance, which means an UE can move to multi-dimensional spaces.

$$S := \{s : s_{\bar{n}} \in \mathbb{U}; \sum_{\bar{n}=1}^{\bar{N}} s_{\bar{n}} \times R_{\bar{n}} \leq f \times \hat{R}_{\bar{n}}\} \tag{23}$$

2) SERVICE RATE FOR UE's

A random UE is deployed to any zone and start its session from the respective BS. An UE can change the zone to one or many times according to the UA. The received nominal service rate Φ for an UE depends upon the current UE location and the availability of bandwidth.[1] The behavior of an UE can show three possible transitions such as $\Phi_{\bar{n}\bar{n}'}$, $\Phi_{\bar{n}-}$, and $\Phi_{\bar{n}0}$. We assume that the number of UE is located in the $\bar{n}$ zone (exponentially distributed) is $\Upsilon_{\bar{n}}$, $\bar{n} = 1, \bar{N}$. In a session, each UE receives the average volume of data which is equivalent to $\frac{1}{\Upsilon_{\bar{n}}}$.

To calculate the nominal service rate, we assume that the nominal bit-rate is provided to an UE by the BS in the $\bar{n}$

[1]The nominal service rate for an UE is based on the channel capacity, resource allocation, and the modulation and coding schemes.

zone is $\hat{\Phi}_{\bar{n},+}$ when there is no deficit in rate. However, as the number of UE increases, the BS starts reducing the bit-rate proportionally. The different bit-rate to the UE is associated with the modulation and coding schemes. The information sent to the UE by the BS is composed of control, data, and re-transmission information. The data information is useful in calculating the service rate, denoted by $1-\mho_{\bar{n}}$, $0<\mho_{\bar{n}}<1$. Hence the nominal service rate to an UE in $\bar{n}$th zone is $\Phi_{\bar{n},+}=\hat{\Phi}_{\bar{n},+}(1-\mho_{\bar{n}})\Upsilon_{\bar{n}}$.

3) UE CALL ARRIVAL APPROACH

We assume that an UE's new calls arrive according to the Poisson arrival process with an arrival rate of ϵ and the exponentially distributed session duration with rate γ. The function $a_{\bar{n}}(s)$ (the simplified form is $a_{\bar{n}}$) shows the session as arrived if the value is 1, end else blocked when the value is 0.

The total area ($Area_{total}$) of the zones is equal to the $\sum_{\bar{n}}^{\bar{N}} Area_{\bar{n}}$. The $Area_{\bar{n}}$ is the area corresponding to the zone $\bar{n}$. The arrival rate for each zone is equal to $\epsilon_{\bar{n}}=\frac{Area_{\bar{n}}}{Area_{total}}\epsilon$, whereas $a_{\bar{n}}$ shows whether a session is arrived in zone k. For example, with two zones ($k=2$) so $\epsilon_1=\frac{Area_1}{Area_{total}}\epsilon$ $\epsilon_2=\frac{Area_2}{Area_{total}}\epsilon$.

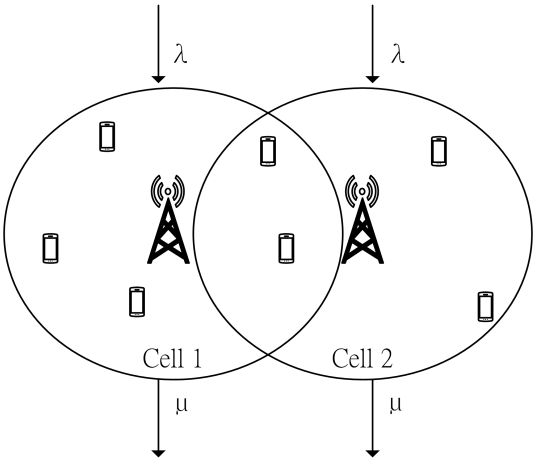

FIGURE 5. Analytical model.

D. STEADY STATE PROBABILITIES OF THE CTMC MODEL

Let X be a CTMC with the state space S. The steady-state probability of a state s, $s\in S$ is $\pi_s=\lim_{t\to\infty}P(X(t)=s)$. Before calculating the transition rate, first we calculate the exiting rate (v) which is the average exits per unit of time. We assume that the state exiting rate from state s is v_s. Thus, the state transition rate ($p'_{ss'}$) of X from state s to state s' is simply the proportion of $p_{ss'}$, i.e. $p'_{ss'}=v_s p_{ss'}$. The local balance equation is $\pi_s p'_{ss'}=\pi_{s'}p'_{s's}$, where $s'=s\pm 1$. Finally, the balance equation for the CTMC is shown in Eq. (24)

$$\pi_{s'}\sum_{k\neq s'}p'_{s'k}=\sum_{k\neq s'}\pi_k p'_{ks'},\ k\in S,\ s\in S,\ s'\in S \quad (24)$$

For simplification, we consider the two-dimensional CTMC model. As shown in Figure 4, a 2-D CTMC model of the birth-death type shows that a network consists of two zones, where each state represents the active number of UEs in the specific zone. The B&D rates of each state depend upon the scheduling policy. In the Figure 4, the rows and columns represent the active number of UEs in a zone. The total number of states in 2-D CTMC is according to $(Q+1)(Q+2)/2$, where Q is the total number of UE in the system. In each state (u,v), u and v represent the active number of UEs in zone 1 and zone 2. The sum of $u+v$ is the total number of active UEs in the system. The transition from state (u,v) to state (u+1,v) shows that a new or handed-over UE is allowed to zone 1, whereas the transition from state (u+1,v) to state (u,v) shows that an UE is either switched off or has departed from zone 1. Similar is the transition from state (u,v) to state (u,v+1) and (u,v+1) to state (u,v) for zone 2. The zone residency time is when an UE stays in a specific zone for a certain amount of continuous time. The arrival rate (λ) and the departure rate (μ) are exponentially distributed as shown in Figure 5. To simplify, we assume that an UE remains in a zone until the session finishes. The B&D rates of the UEs in the model are state-dependent. Let $\pi_{u,v}$ denote the steady-state probability of state (u,v) in 2-D CTMC model. The balance equations can be obtained by equating the transition probabilities of each state. From the 2-D CTMC state diagram we can have the rate of transition for each state. For the state (u,v), the transition rate for "into the state" is $(v+1)\lambda\times\pi_{u-1,v+1}+a_1\epsilon_1\times\pi_{u-1,v}+v\gamma\times\pi_{u,v-1}+(u+1)\mu\times\pi_{u+1,v-1}$ and for "onto the state" is $[u\mu+2\gamma+a_2\epsilon_2+v\lambda]\pi_{u,v}$. These two rates must be equal. The 2-D CTMC steady state probability π=[$\pi_0,\pi_1,\ldots,\pi_Q$] for Q active UE, and P is the transition probability matrix such as $\pi P=0$. The normalization equation is as shown in Eq. 25. The balance equations can be solved by using direct substitutions. We have used MATLAB to solve these equations and derive steady-state probabilities of the 2-D CTMC model for the u+v active UEs in the network.

$$\sum_{u=0}^{Q}\sum_{v=0}^{Q-u}\pi_{u,v}=1 \quad (25)$$

For example, when $Q=3$, the number of states in the 2-D CTMC model is (3+1)(3+2)/2 = 10. According to Figure 6, the transition matrix is in Eq. 26, as shown at the bottom of the next page.

Note that, based on modulation and coding schemes (MCS) each cell can further be divided. According to this the channel conditions and throughput also vary. In our analytical program, we have considered only one MCS value which means an UE is receiving uniform radio conditions throughout the zone. Each cell is represented by a zone. For simplification, we have considered only two zones and one MCS value for the analytical model providing a bandwidth of 3 and 5 MHz.

VI. SIMULATION RESULTS

In this section, we develop a simulator in MATLAB and validate the effectiveness of our scheme in the 5G HetNet

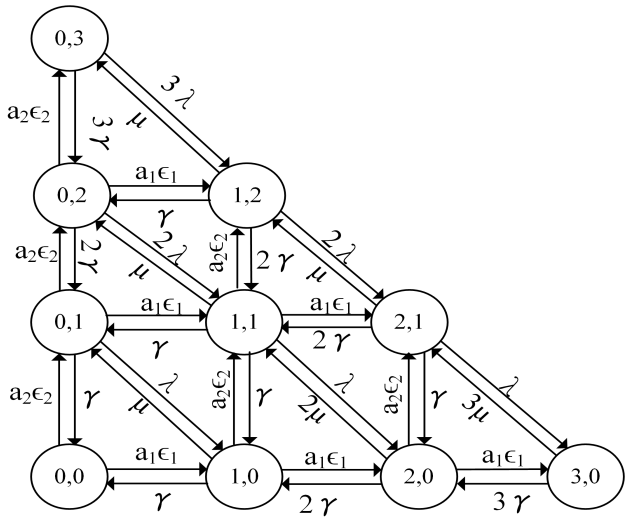

FIGURE 6. Two-Dimensional CTMC model, when Q or $(u + v) = 3$.

environment. In this study, we have considered an MBS and six small BSs, which are connected to distinct CCs of 1.4, 3, 5, 10, 15, and 20 MHz of bandwidth. Channel qualities are adopted in the simulation, as shown in Table 3. In the current study, for simplification to validate the performance, we considered CA of maximum 2 CCs; which can be easily extended to accommodate a greater number of CCs. In addition, we have deployed Poisson packet data traffic with pseudo-random integral packet inter-arrival rate in the range between 0 and a maximum of 50 ms. In a simulation, each UE is connected to a general traffic, receiving data packets connected to three applications/flows containing two GBR and one non-GBR applications similar to [9]. To compare our simulation results we have considered different scheduling schemes, that are first-come-first-serve (FCFS) [26], PF [27], Modified Largest Weighted Delay First (MLWDF) [28] and Serving-Ratio (SR) [29]. In order to check the effectiveness of our scheme, we under-load and over-load the network while varying the number of UEs from 200 to 1000. The simulation parameters are shown in Table 4. Note that each simulation result has been averaged by at least 1000 experiments.

We consider five performance metrics: (i) Wake-up Ratio: The ratio of wake-up sub-frames over the total execution sub-frames of an UE; (ii) Average Delay: It is the sum of delays introduced by the traffic model and the DRX. Each packet has an inter-packet arrival time, which follows an exponential distribution with some means that cause some delay. And the extra delay is due to the active mode of DRX mechanism by some packet scheduling mechanism by the BS; (iii) Average Network Throughput (Mbps) (iv) Number of UEs: It is the number of UEs in the corresponding cell which have been allocated some number of RBs in a time frame, as in over-load scenario some UEs may not be allocated RBs; and (v) Data Dropped Rate: It is the ratio of number of the packets failed to receive over the total number of packets of the UE. To compare, we have used previous schemes such as FCFS, PF, MLWDF, and SR packet data traffic schemes.

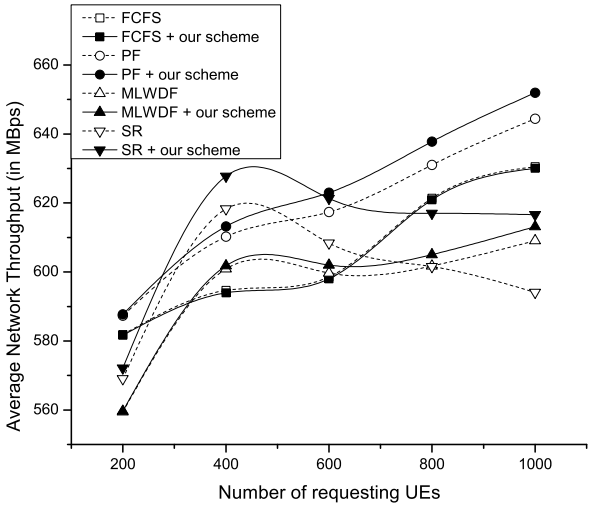

FIGURE 7. The average network throughput (in Mbps) while varying the number of UEs.

First, we investigate how the number of requesting UEs affect the average Network Throughput. In general, while increasing the number of UEs, throughput also increases proportionally [19]. As shown in Figure 7, the average network throughput (in Mbps) increases while employing our scheme with data packet scheduling. In our simulation,

$$\mathbf{P} = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ -a_1\epsilon_1 & a_1\epsilon_1 & 0 & 0 & a_2\epsilon_2 & 0 & 0 & 0 & 0 & 0 \\ \gamma & -(a_1\epsilon_1+\gamma) & a_1\epsilon_1 & 0 & \mu & a_2\epsilon_2 & 0 & 0 & 0 & 0 \\ 0 & 2\gamma & -(a_1\epsilon_1+2\gamma) & a_1\epsilon_1 & 0 & 2\mu & a_2\epsilon_2 & 0 & 0 & 0 \\ 0 & 0 & 3\gamma & 0 & 0 & 0 & 3\mu & 0 & 0 & 0 \\ \gamma & \lambda & 0 & 0 & 0 & a_1\epsilon_1 & 0 & a_2\epsilon_2 & 0 & 0 \\ 0 & \gamma & \lambda & 0 & \gamma & 0 & a_1\epsilon_1 & \mu & a_2\epsilon_2 & 0 \\ 0 & 0 & \gamma & \lambda & 0 & 2\gamma & 0 & 0 & 2\mu & 0 \\ 0 & 0 & 0 & 0 & 0 & 2\lambda & 0 & 0 & a_1\epsilon_1 & a_2\epsilon_2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 2\lambda & \gamma & 0 & \mu \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 3\gamma & 3\lambda & 0 \end{pmatrix} \quad (26)$$

TABLE 3. LTE-A standards modulation and coding schemes [19].

Radio Bearer index	Modulation	Channel Coding Rate	Bearer Efficiency (bits/symbol)	SINR (dB)
1	QPSK	0.0761719	0.1523	-6.936
2	QPSK	0.117188	0.2344	-5.147
3	QPSK	0.188477	0.377	-3.180
4	QPSK	0.300781	0.6016	-1.253
5	QPSK	0.438477	0.877	0.761
6	QPSK	0.587891	1.1758	2.699
7	16QAM	0.369141	1.4766	4.694
8	16QAM	0.478516	1.9141	6.525
9	16QAM	0.601563	2.4063	8.573
10	64QAM	0.455078	2.7305	10.366
11	64QAM	0.553711	3.3223	12.289
12	64QAM	0.650391	3.9023	14.173
13	64QAM	0.753906	4.5234	15.888
14	64QAM	0.852539	5.1152	17.814
15	64QAM	0.925781	5.5547	19.829

TABLE 4. Simulation parameters.

Parameters	Values
Number of UEs	200 $\sim$ 1000
Thermal noise level	-104 dBm/Hz
MBS power level	24 $\sim$ 44 dBm
Pico base station power level	30 $\sim$ 35 dBm
Femto base station power level	20 $\sim$ 24 dBm
Pathloss (PL) parameter	For MBS α=128 and β=37.6
PL=α+$\beta \times log_{10}$(d)[dB] d in meters	For PBS and FBS α=38 and β=30
Transmission time interval	1ms

we have considered an over-load scenario above 75%. In our case, our total RBs is 441. At 400 number of UEs, there is a change in all the curves, as from here our scheme effectively offloads data to other SBSs. The PF shows better performance since it gives higher priority to that UE, whose rate performance is better than others. The MLWDF also shows a similar trend but with lesser average network throughput. Since MLWDF loses more number of packets as mentioned in section III-B, the SR shows a similar curve which is a very similar trend as shown in paper [29]. Once the network is in an overload scenario while SR considers the remaining serving ratio, it determines the allocation order. It also maintains the lower average delay (which is evident from Figure 9) but SR + our scheme flattens in the over-load scenario since it gives higher priority to the UEs having less DAD in the DCPF parameter.

Secondly, we investigate the number of serving UEs with the number of requesting UEs as shown in Figure 8. By implementing our scheme with the data packet scheduling, all of them show a similar increase in serving UEs. The PF and MLWDF show less number of served UEs, as they give higher priority to the higher data rate UEs. While introducing our scheme both show a significant change as more number of UEs get connected to it. In our scheme, the RARF enhances those UEs that are at the edge, since they show a significant change in allocation while shifting from over-loaded BSs to the under-loaded BSs. That is why they both show a

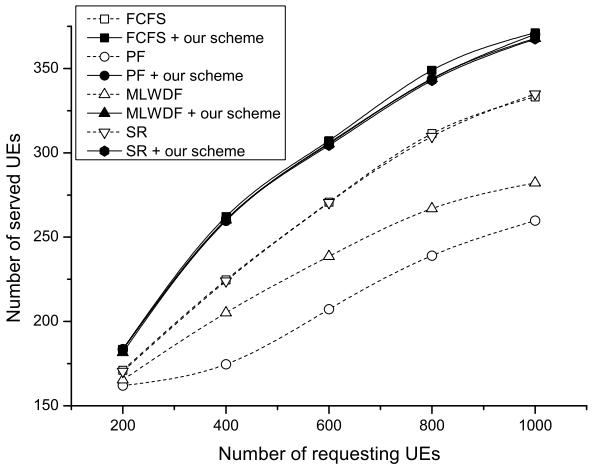

FIGURE 8. Assigned UEs to the network while varying the number of UEs.

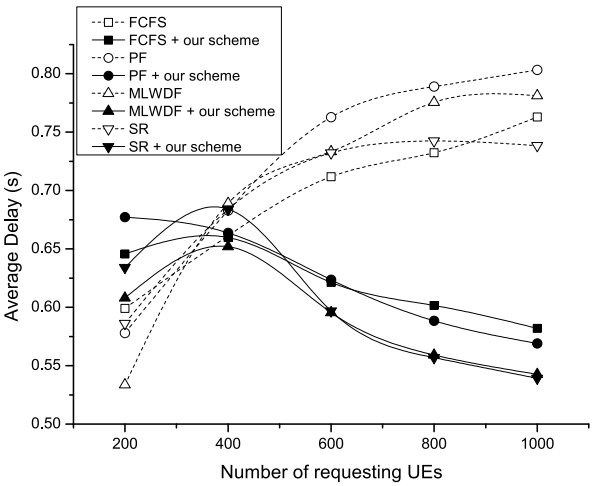

FIGURE 9. Average delay while varying the number of UEs.

significant change in the served UEs. Whereas, the FCFS and SR also improve the number of served UEs. Next, we have investigated the Average delay as shown in Figure 9. Our scheme has significantly reduced the average delay, as the

first phase of our proposed scheme guarantees the QoS of an UE, which can be witnessed in results. In our scheme, the reason of increase, is as the network is getting saturated and it becomes difficult to serve all UEs' packets under the consideration of packet delay budgets. Once the overloading phase triggered (near to 400 number of requesting UEs) according to the Eq. (21) and (22). Our scheme start selecting those UEs whose DCPF value is low. DCPF is the product of delay and power consumption. Due to this there will be higher priority to lower delay UEs. Lower delay is also one of the favorable conditions for the UEs. That is why the execution of our scheme significantly reduces the average delay.

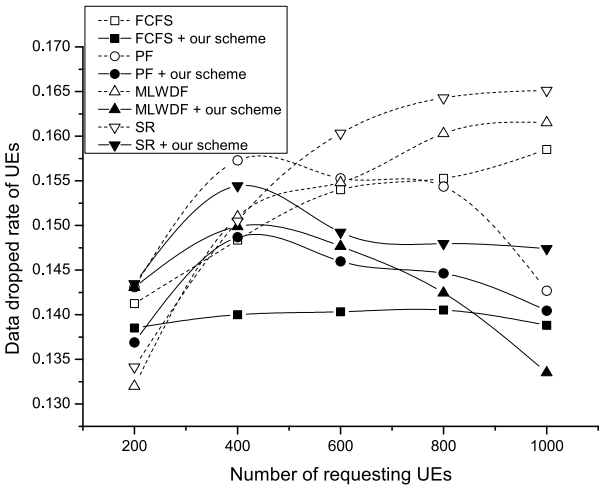

FIGURE 10. The data dropped rate of UE.

Additionally, we explore the data dropout rate, as depicted in Figure 10. Our scheme has shown a reduced dropped rate while increasing the number of requesting UEs. As it is known that an over-loaded BS shows higher data dropped rate whereas, in an under-loaded BS always shows a reduced dropped rate. Since our scheme shifts UEs from over-loaded BSs to under-loaded, it means that the dropped rate has to be reduced which is evident from the results.

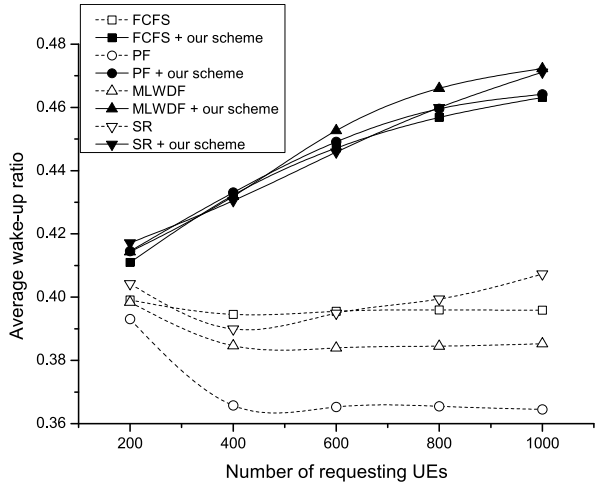

FIGURE 11. The DRX wake-up ratio while varying the number of UEs.

Finally, we investigate the average wake-up ratio of different data packet scheduling with and without our scheme as shown in Figure 11. The increased wake-up ratio shows the overhead cost introduced by the increased number of serving UEs. The average increased number of UEs (For example (FCFS+our scheme) - FCFS) for PF, MLWDF, FCFS and SR are 84, 60, 32, and 29 respectively. Similarly, the average increased wake-up ratio of UEs for PF, MLWDF, FCFS and SR are 0.07, 0.06, 0.0458, and 0.0457 respectively. From this we can say that the increased wake-up ratio is due to the overhead cost introduced by the increased number of UEs. The PF and MLWDF shows the least wake-up ratio as less number of UEs connected to them, whereas after introducing our scheme to PF and MLWDF, there is an increase in number of UE, which results in higher wake-up ratio.

VII. LIMITATIONS OF THE CURRENT MODEL

Following studies [30], [31], [32], [33] contribute valuable insights and their limitations into queuing analysis and mathematical modeling for mobile communication networks. Studies [30], [31] introduces a mathematical model for the queueing analysis of mobile communication network cells, considering the operation with mobile users. The model computes the key performance measures based on cell bandwidth and zone division. User activation processes in different zones are defined using a marked Markovian arrival process. A noted limitation is the common representation of zones as concentric rings, overlooking the general variability due to interference with surrounding Base Stations. To simplify, the authors focuses solely on a single user class. In contrast, another study [32] employed multi-dimensional model, aligned with a Markov process, exhibits insensitivity to the distribution of UE session duration, as the state probabilities exclusively rely on the mean service time [34]. However, this model imposes constraints on inter-dimensional dependence, limiting the number of UEs for a specific zone to an upper bound. In each region, UEs exhibit varying channel qualities within the cell, potentially encompassing a considerable range of values. To simplify analysis, the study assumes that an active UE's session occurs in one of these zones with a specified probability. Another analysis [33] acknowledges that the total service time for an UE in a cell involves visiting a random number of zones during its sojourn. This results in the total service time being a combination of a random number of exponential variables, with parameters dependent on the visited zone. Study suggests the use of phase-type (PH) distribution, known for describing total service times in cells. However, the challenge arises as the traditional PH distribution is designed for homogeneous customers, whereas the model in consideration involves heterogeneous customers. In the proposed model, the total service time can conclude either successfully or non-successfully, with an UE departing from the cell before service completion—a differentiation not captured by the traditional PH distribution.

In conclusion, the varied perspectives on queuing analysis and mathematical modeling presented in these studies provide valuable insights. While one focuses on performance measures and zone representations, the other explores multi-dimensional considerations, emphasizing the limitations of current model in heterogeneous users. These findings helps to simplify the complex nature of model and also guide the refinement of my research, contributing to a more adaptable mathematical model for mobile communication networks.

VIII. CONCLUSION AND FUTURE WORK

In this study, we have addressed the offloading mechanism by effective UA while considering the UE's power, and guaranteed QoS in dense HetNet environment. We have proposed a meta-heuristic three-phase scheme that optimizes the performance of both UEs and the network. In the first phase, DRX optimization problem is addressed that configures the DRX parameters based on the QoS such as traffic bit-rate, packet delay budget, and packet loss rate while saving power consumption of UEs. To evaluate the performance of UEs in terms of power and delay, we consider the DCPF, and for the BSs load, we calculate the relative change of allocated RBs to the current and neighboring BSs, i.e., RARF. The DCPF and RARF helps in optimizing the network throughput and UE's power performance. The second phase estimates the DCPF and RARF performance factors, whereas the third phase triggers the offloading. Extensive simulation results show that our proposed scheme improves the system throughput and decreases the average delay of an UE significantly. Our scheme also increases the number of served UEs in the network. The increase in the wake-up ratio is due to the overhead cost of the increased number of serving UEs. We have also analyzed the analytical results while employing the 2-D CTMC and semi-Markov model.

In future research, the objective is to enhance the mathematical model for heterogeneous mobile communication networks by incorporating various user classes, including real-time versus non-real-time users, static and mobile users, diverse mobility profiles, and varying QoS demands of different applications and services. This effort seeks to provide practical solutions for optimizing system performance across a wide range of user scenarios. Additionally, another research direction explores a scheme grounded in reinforcement learning. This approach aims to optimize DRX configuration parameters and further refine network optimization strategies.

REFERENCES

[1] S. Parkvall, E. Dahlman, A. Furuskar, and M. Frenne, "NR: The new 5G radio access technology," *IEEE Commun. Standards Mag.*, vol. 1, no. 4, pp. 24–30, Dec. 2017.

[2] C. J. Zhang, J. Ma, G. Y. Li, Y. Kishiyama, S. Parkvall, G. Liu, and Y.-H. Kim, "Key technology for 5G new radio," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 10–11, Mar. 2018.

[3] S. Parkvall, Y. Blankenship, R. Blasco, E. Dahlman, G. Fodor, S. Grant, E. Stare, and M. Stattin, "5G NR release 16: Start of the 5G evolution," *IEEE Commun. Standards Mag.*, vol. 4, no. 4, pp. 56–63, Dec. 2020.

[4] E. Rastogi, M. K. Maheshwari, A. Roy, N. Saxena, and D. R. Shin, "Machine learning-based DRX mechanism in NR-unlicensed," *IEEE Wireless Commun. Lett.*, vol. 11, no. 5, pp. 1052–1056, May 2022.

[5] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.

[6] *Evolved Universal Terrestrial Radio Access (E-UTRA); User Equipment (UE) Procedures in Idle Mode (Release 8)*, document TS 36.304, Version 8. 0.0, 3GPP, Dec. 2008.

[7] B. B. Kumar, L. Sharma, and S.-L. Wu, "Online distributed user association for heterogeneous radio access network," *Sensors*, vol. 19, no. 6, p. 1412, Mar. 2019.

[8] A. Mesodiakaki, F. Adelantado, L. Alonso, and C. Verikoukis, "Energy-efficient user association in cognitive heterogeneous networks," *IEEE Commun. Mag.*, vol. 52, no. 7, pp. 22–29, Jul. 2014.

[9] J.-M. Liang, J.-J. Chen, H.-H. Cheng, and Y.-C. Tseng, "An energy-efficient sleep scheduling with QoS consideration in 3GPP LTE-advanced networks for Internet of Things," *IEEE J. Emerg. Sel. Topics Circuits Syst.*, vol. 3, no. 1, pp. 13–22, Mar. 2013.

[10] M. K. Maheshwari, M. Agiwal, N. Saxena, and A. Roy, "Directional discontinuous reception (DDRX) for mmWave enabled 5G communications," *IEEE Trans. Mobile Comput.*, vol. 18, no. 10, pp. 2330–2343, Oct. 2019.

[11] L. Sharma, B. B. Kumar, and S.-L. Wu, "Performance analysis and adaptive DRX scheme for dual connectivity," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10289–10304, Dec. 2019.

[12] J.-M. Liang, P.-Y. Chang, J.-J. Chen, C.-F. Huang, and Y.-C. Tseng, "Energy-efficient DRX scheduling for D2D communication in 5G networks," *J. Netw. Comput. Appl.*, vol. 116, pp. 53–64, Aug. 2018.

[13] R. Sharma, N. Kumar, N. B. Gowda, and T. Srinivas, "Packet scheduling scheme to guarantee QoS in Internet of Things," *Wireless Pers. Commun.*, vol. 100, no. 2, pp. 557–569, May 2018.

[14] T. K. Vu, M. Bennis, S. Samarakoon, M. Debbah, and M. Latva-aho, "Joint load balancing and interference mitigation in 5G heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 6032–6046, Sep. 2017.

[15] W. C. Ao and K. Psounis, "Approximation algorithms for online user association in multi-tier multi-cell mobile networks," *IEEE/ACM Trans. Netw.*, vol. 25, no. 4, pp. 2361–2374, Aug. 2017.

[16] X. Ge, X. Li, H. Jin, J. Cheng, and V. C. M. Leung, "Joint user association and user scheduling for load balancing in heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 5, pp. 3211–3225, May 2018.

[17] H. U. Sokun, R. H. Gohary, and H. Yanikomeroglu, "A novel approach for QoS-aware joint user association, resource block and discrete power allocation in HetNets," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7603–7618, Nov. 2017.

[18] X. Luo, "Delay-oriented QoS-aware user association and resource allocation in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1809–1822, Mar. 2017.

[19] Y.-C. Wang and C.-C. Huang, "Efficient management of interference and power by jointly configuring ABS and DRX in LTE—A HetNets," *Comput. Netw.*, vol. 150, pp. 15–27, Feb. 2019.

[20] M. R. G. Aghdam, B. Rahmani, and R. Abdolee, "Traffic-based adjustable discontinuous reception mechanism with bounded delay," in *Proc. 16th Annu. Conf. Wireless On-Demand Netw. Syst. Services Conf. (WONS)*, Mar. 2021, pp. 1–6.

[21] R. Mudumbai, S. K. Singh, and U. Madhow, "Medium access control for 60 GHz outdoor mesh networks with highly directional links," in *Proc. IEEE INFOCOM*, Apr. 2009, pp. 2871–2875.

[22] H. Zhang, S. Huang, C. Jiang, K. Long, V. C. M. Leung, and H. V. Poor, "Energy efficient user association and power allocation in millimeter-wave-based ultra dense networks with energy harvesting base stations," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 9, pp. 1936–1947, Sep. 2017.

[23] S. Stanczak, M. Wiczanowski, and H. Boche, *Fundamentals of Resource Allocation in Wireless Networks: Theory and Algorithms*. Berlin, Germany: Springer, 2009.

[24] M. M. Hasan, S. Kwon, and J.-H. Na, "Adaptive mobility load balancing algorithm for LTE small-cell networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 4, pp. 2205–2217, Apr. 2018.

[25] B. Sas, E. Bernal-Mor, K. Spaey, V. Pla, C. Blondia, and J. Martinez-Bauset, "Modelling the time-varying cell capacity in LTE networks," *Telecommun. Syst.*, vol. 55, no. 2, pp. 299–313, Feb. 2014.

[26] N. Becker and M. Fidler, "A non-stationary service curve model for estimation of cellular sleep scheduling," *IEEE Trans. Mobile Comput.*, vol. 18, no. 1, pp. 28–41, Jan. 2019.

[27] S. Dawaliby, A. Bradai, Y. Pousset, and C. Chatellier, "Joint energy and QoS-aware memetic-based scheduling for M2M communications in LTE-M," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 3, no. 3, pp. 217–229, Jun. 2019.

[28] V. K. Shrivastava, P. Makhija, and R. Raj, "Joint optimization of energy efficiency and scheduling strategies for side-link relay system," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2017, pp. 1–6.

[29] C.-K. Hsu, J.-M. Liang, K.-R. Wu, J.-J. Chen, and Y.-C. Tseng, "Enhanced scheduling schemes with energy conservation for dynamic point selection in cloud radio access networks," *Wireless Netw.*, vol. 26, no. 2, pp. 1519–1534, Feb. 2020.

[30] C. Kim, S. A. Dudin, O. S. Dudina, and A. N. Dudin, "Mathematical models for the operation of a cell with bandwidth sharing and moving users," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 744–755, Feb. 2020.

[31] C. Kim, A. Dudin, S. Dudin, and O. Dudina, "Mathematical model of operation of a cell of a mobile communication network with adaptive modulation schemes and handover of mobile users," *IEEE Access*, vol. 9, pp. 106933–106946, 2021.

[32] O. Adamuz-Hinojosa, P. Ameigeiras, P. Muñoz, and J. M. Lopez-Soler, "Analytical model for the UE blocking probability in an OFDMA cell providing GBR slices," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Mar. 2021, pp. 1–7.

[33] S. Dudin and C. Kim, "Analysis of multi-server queue with spatial generation and location-dependent service rate of customers as a cell operation model," *IEEE Trans. Commun.*, vol. 65, no. 10, pp. 4325–4333, Oct. 2017.

[34] V. B. Iversen, "Teletraffic engineering and network planning," Dept. Photon. Eng., Tech. Univ. Denmark, Lyngby, Denmark, Tech. Rep., 2010, pp. 1–399.