QoS-aware Resource Allocation for Live Streaming in Edge-Clouds Aided HetNets Using Stochastic Network Calculus

Abbas Mirzaei (mirzaei.iaut@gmail.com)
Islamic Azad University

Research

Keywords: Resource Allocation, Quality of Service, Multi-layer Optimization, Data Caching, Mobile Edge Computing (MEC)

Posted Date: June 28th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-632457/v1

Version of Record: A version of this preprint was published at Reliability Engineering & System Safety on April 1st, 2022. See the published version at https://doi.org/10.1016/j.ress.2021.108272.

QoS-aware Resource Allocation for Live Streaming in Edge-Clouds Aided HetNets Using Stochastic Network Calculus

Abbas Mirzaei [1*], *Senior Member, IEEE*
[1]Department of Computer Engineering, Ardabil Branch, Islamic Azad University, Ardabil, Iran
[*] Corresponding author's email address: *a.mirzaei@ieee.org*

Abstract
Mobile edge computing (MEC) is a key feature of next generation mobile networks aimed at providing a variety of services for different applications by performing related processing tasks closer to the user equipment. Edge clouds can be installed as an interface between the cellular networks and the core to provide the required services based on the known concept of the MEC networks. Nonetheless, the problem of green networking will be of great importance in such networks. This paper presents an energy-efficient stochastic network calculus (SNC) framework to control MEC data flows. In accordance with the entrance processes of different QoS-class data flows, closed-form problems were formulated to determine the correlation between resource utilization and the violation probability of each data flow. Also, in the access layer, this paper proposes a dynamic user association and resource allocation approach which maximizes the overall energy efficiency of cache-enabled cellular networks in addition to provide the superior fairness level for UEs. In this energy-cooperative approach, the power can be shared among the cells using a grid network. This model also performs routing in the multi-hop backhaul to efficiently use the existing infrastructure of small cell networks for simultaneous dual-hop transmissions. The simulation results exhibit that the proposed approach can effectively increase the user throughput and the total power efficiency while guaranteeing the acceptable fairness level for uniform and hotspot UE distribution models. It also proved that the energy utilization index and the system data rate can be significantly improved.

Index Terms: Resource Allocation, Quality of Service, Multi-layer Optimization, Data Caching, Mobile Edge Computing (MEC).

I. INTRODUCTION

Nowadays, the growing use of digital mobile devices has sharply increased the data flows of different commercial applications. Network delays caused by the growing quantity of UEs can challenge a network's capacity for high data rate services [1]. To cope with such problems, caching can be used as an appropriate method for decreasing delay and system overload via offloading contents from core layer to other layers of the system or even to UEs [2].

Mobile edge computing networks supply closer data-caching and computing services for UEs to decrease the end-to-end latency and power utilization. In other words, storage and computing are executed via edge nodes at a shorter distance to the UEs [3]. The mobile edge computing network can usually include different QoS-class data flows. The software-defined networking (SDN) technology can also embedded into the mobile edge computing to adaptively control network traffic, which results in the random-routing of data streams. Therefore, to develop an appropriate analytical framework for various applications in the mobile edge computing network, the concept of random routing should be integrated into the analytical framework.

The end-to-end performance evaluation of various data streams were analyzed in different scenarios in [4], which the authors applied dynamic queue theory in these scenarios. It is considerable that the results of utilizing queue theory were equal to the average results of analysis in steady states. Stochastic network calculus can be used as a network performance analysis software plane to alter complicated non-linear frameworks to simple linear frameworks through min/max-plus method. In addition, the stochastic network calculus is significantly simple and dynamic; hence, it is proper for the analysis of various scenarios. This paper applied stochastic network calculus to control the end-to-end functionality of QoS-aware data streams of cache-enabled MEC networks through random routing. SNC was employed in [5] to analyze end-to-end latency considering interference-coordination in wireless distributed networks. In [6], the researchers analyzed the performance of multi-server networks and, in particular, the effect that the number of servers would have on network functionality. Using stochastic network calculus, the researchers in [7] indicated the effect of carrier allocation and signaling on the latency of packets in multicarrier networks considering the total capacity of the transmission system as a constant value. Some other studies have employed the moment-generating function (MGF) as a type of stochastic network calculus approach to analyze the system's functionality [8].

For the access layer, fifth-generation networks are predicted to have higher energy efficiency, which is one of the prominent characteristics of NGMNs. Using renewable energy sources to supply the energy of base stations can be considered as a viable solution for reducing energy consumption and solving environmental problems such as high carbon emissions. Nevertheless, the energy acquired from renewable energy sources is not constant, and this has a great effect on the quality of service in cache-enabled cluster networks with renewable power utilization.

According to the previous studies, some frameworks have been developed for dynamic power optimization in cache-enable systems and power sharing, respectively [9]–[11]. Nevertheless, robust power optimization has not been performed in NOMA networks which have both of these capabilities simultaneously, and further studies are required. In this regard, the current paper proposes a joint user association and resource allocation framework in a cache-enabled heterogeneous network with the capability of power-cooperation, aiming to maximize the network's throughput and minimize power utilization. Also, according to the research literature, stochastic network calculus has never been employed to analyze end-to-end performance and to model a framework for different applications of the mobile edge computing network through random routing. Thus, the results of the current study can concurrently be applied for efficient routing and network resource allocation and scheduling.

The remainder of this paper is structured as follows. In Section II, we present the mathematical model and its assumptions including the caching strategy, energy model and the problem formulation. In section III we present our resource allocation and user association approaches considering QoS and energy efficiency constraints. Section IV is dedicated to data-flow in mobile edge computing networks. In this section we tried to suggest an end-to-end performance analysis using stochastic network calculus. Section 5 presented the numerical evaluation and details of the parameters that have been applied in the scenarios. This section considered the effect of different user demands and user distribution patterns on the achievable power savings for backhaul and access network, data rate and fairness index. Finally, Section 6 concludes the paper and lists ideas for future work.

II. METHODS/EXPERIMENTAL

The system model is a cache-enabled cooperative heterogeneous network including a macro base station (*MBS*) and numerous small cells, so that each base station is empowered with caching capability. Indicators U and B represent the set of UEs and the set of base stations respectively. L_S and L_M indicate respectively the cache size of each small cell and macro cell. In this scenario, the energy of each base station is provided by both renewable resources and smart grids in which base stations can share their power through a grid network. Based on this scenario, the configuration of the caching-enabled HetNet exhibited in Fig. 1.

A. Caching Strategy

In this caching strategy, it's assumed that there is a limited number of data frame demonstrated as $\mathcal{F} = \{\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_f, \dots, \mathcal{F}_F\}$, in which, $\mathcal{F}_f$ and F denote the f-th data frame and the number of frames respectively. Each data frame has specific size and the probability that a UE requests data frame f is represented as

$$p_f\big(0 \le p_f \le 1\big), \qquad which, \sum_{f=1}^{F} p_f \le L_i, \ \ \forall f \in \mathcal{F}, \tag{1}$$

In this paper we applied the stochastic caching strategy. The probability that cell i caches a specific data frame f is $0 \le q_{f_i} \le 1$ in which L_i shows the cache size of cell i. Note that $\{q_{f_i}\}$ for cell i should meet the following constraint [12]:

$$\sum_{f=1}^{F} q_{f_i} \le L_i, \ \ \forall i \in \mathcal{B}, f \in \mathcal{F}, \tag{2}$$

If the serving cell is macro cell, we have $L_i = L_M$, and for small cells $L_i = L_S$.

B. Energy Model

The main energy resources for the base stations are renewable energy and smart-grid. In each time slot, the transmission power of cell i denoted as $P_i (i \in B)$, the smart-grid power used by cell i is G_i, and the power harvested by cell i is expressed by E_i. In the cooperation mode, the transferred power from base station i to base station i' shown by $\varepsilon_{ii'}$, in which $\beta \in [0,1]$ represents the power-transfer efficiency index among cells. Hence, $(1-\beta)$ indicates the power loss percentage during the power sharing process. The transmission power at the i-th base station must meet the following constraint.

$$P_i < G_i + E_i + \beta \sum_{i' \in \mathcal{B}, i' \neq i} \varepsilon_{i'i} - \sum_{i' \in \mathcal{B}, i' \neq i} \varepsilon_{ii'}. \tag{3}$$

Based on this formulation, the total smart-grid power utilization is directly dependent on the harvested power, shared power and the transmission power of base stations.

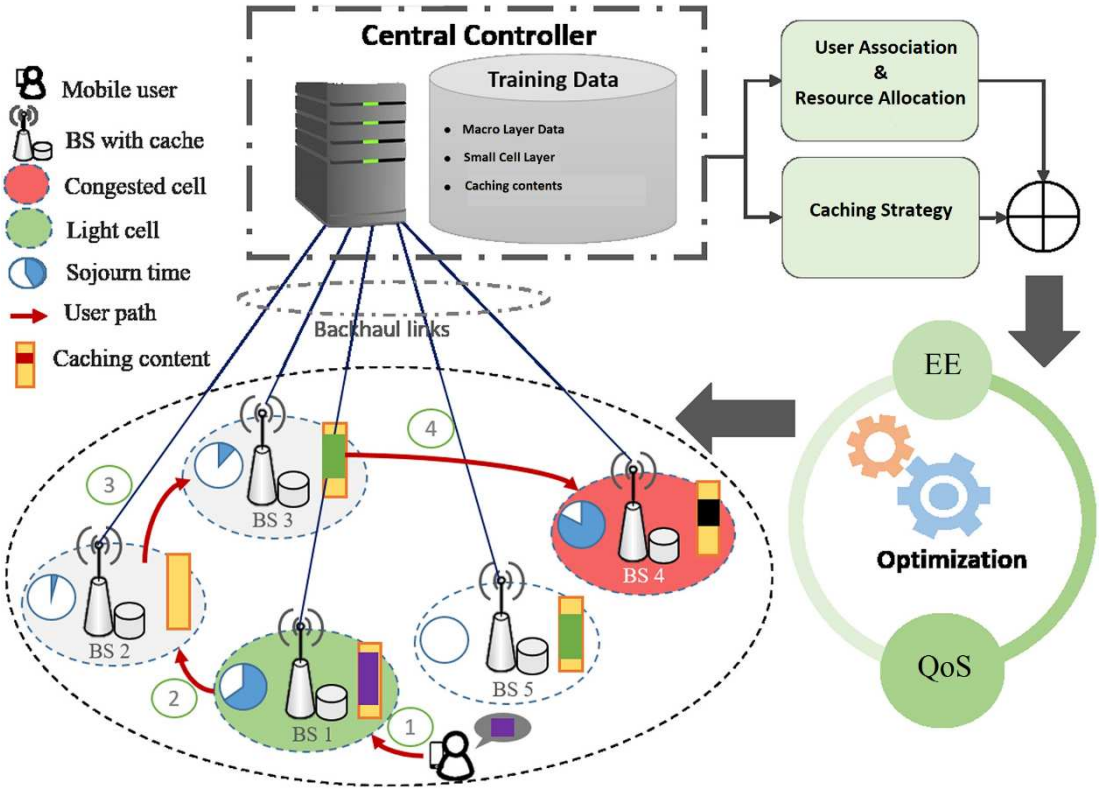

Fig. 1. Caching-enabled HetNet Configuration

C. Problem Formulation

In order to guarantee fairness, we considered the proportionally throughput balancing among users [13] in which the sum of the logarithmic utilities over all user equipment should be considered to decrease the data rate unbalancing. In this formulation $x_{ij}(i \in \mathcal{B}, \mathrm{j} \in \mathcal{U})$ demonstrates the binary user association index, for example, if $x_{ij} = 1$ user j is associated to base station i and otherwise it will be equal to zero. So, $k_i = \sum_{\mathrm{j}\in\mathcal{U}} x_{ij}$ denotes the number of UEs served by cell i, and $\left(\sum_{f=1}^{F} p_f q_{f_i}\right)^{k_i}$ shows the probability that k_i associated users are able to be served by cell i which the base station caches their requested data frames. If $x_{ij} = 1$, the utility of the j-th UE is expressed as $\mu_{ij} = \log(R_{ij})$ with the throughput R_{ij}. Here, the throughput R_{ij} (in bits/s) of the j-th UE is obtainable as.

$$R_{ij} = \left(\sum_{f=1}^{F} p_f q_{f_i}\right)^{k_i} \frac{\mathcal{B}}{\sum_{j\in u} x_{ij}} \log(1+\gamma_{ij}) \tag{4}$$

The signal to interference-noise ratio is obtained as (5)

$$\gamma_{ij} = \frac{P_i h_{ij}}{\sum_{i'\in\mathcal{B}, i'\neq i} P_{i'} h_{i'j} + \sigma^2} \tag{5}$$

In which, the bandwidth is shown as B, h_{ij} demonstrated the channel gain between user j and the associated base station i, $h_{i'j}$ exhibits the interfering channel gain between user j and base station i', and σ^2 indicates the noise power. We can claim that the performance level is related to carrier quality and hit probability, which the hit probability is a very effective parameter on the throughput in the energy cooperative HetNets, as is shown by (4). The target is to maximize the system performance while decreasing the total grid power utilization. The goal function can be formulated as

$$\textbf{P1:} \quad \max_{q,x,P,\varepsilon,G} \sum_{i\epsilon\mathcal{B}} \sum_{j\epsilon u} x_{ij}\mu_{ij} - \eta \sum_{i\in\mathcal{B}} G_i \tag{6}$$

s. t.

$$C1: \sum_{i\in\mathcal{B}} x_{ij}\gamma_{ij} \geq \gamma_{min}, \forall j \in u,$$

$$C2: \sum_{i\in\mathcal{B}} x_{jm} = 1, \forall j \in u,$$

$$C3: P_i < G_i + \beta \sum_{i'\in\mathcal{B}, i'\neq \mathrm{i}} \varepsilon_{i'i} - \sum_{i'\in\mathcal{B}, i'\neq \mathrm{i}} \varepsilon_{ii'} + E_i, \forall i \in \mathcal{B},$$

$$\mathrm{C4}: \sum_{f=1}^{F} qf_i \leq L_i, \forall i \in \mathcal{B}, f \in \mathcal{F},$$

$$\mathrm{C5}: 0 \leq q_{f_i} \leq 1, \forall f \in \mathcal{F}, \forall \mathrm{i} \in \mathcal{B},$$

$$\mathrm{C6}: x_{ij} \in \{0,1\}, \forall \mathrm{i}, \forall \mathrm{j} \in u,$$

$$\mathrm{C7}: G_i \geq 0, \varepsilon_{ii'} \geq 0, \forall \mathrm{i} \in \mathcal{B},$$

$$\mathrm{C8}: 0 \leq P_i \leq P_{max}^{i}, \forall \mathrm{i} \in \mathcal{B},$$

In this formulation, $\mathbf{q}=[q_{f_i}]$, $\mathbf{X}=[x_{ij}]$, $\mathbf{P}=[P_i]$, $\boldsymbol{\mathcal{E}}=[\boldsymbol{\varepsilon}_{\mathrm{i}i'}]$, $\mathbf{G}=[G_{\mathrm{i}}]$, η indicates a weighted variable as energy efficiency index, and γ_{min} indicates the minimum signal to noise/interference needed by a user equipment. Also, C1 represents the throughput constraints; C2 and C6 guarantee that a user is not able to be connected to multiple base stations at the same time; C3 is relevant to the power utilization limitation; C4 and C5 represent stochastic caching rules, which were applied in (2); constraint C7 mentions that the applied power and the shared power should be positive values, and C8 indicates the upper bound of allowed transmit power. Note that in this scenario, the multi-hop backhauling configuration of the mobile edge computing HetNet demonstrated in Figure 2.

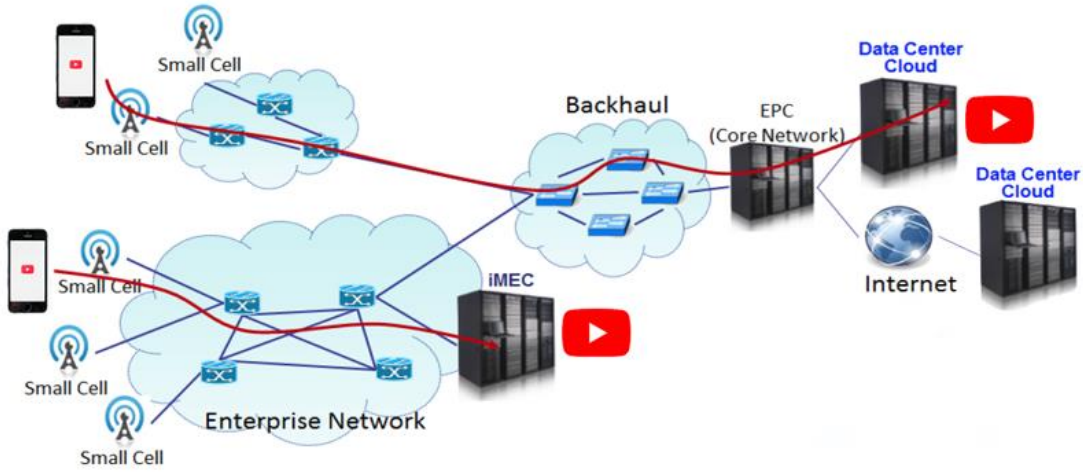

Fig. 2: Multi-hop backhauling strategy in mobile edge computing HetNets

III. Resource Allocation And User Association In Cache-Enabled HetNets

In this part we present our resource allocation and user association approach which can be considered as a solution for problem P1. Considering $k_i = \sum_{j\in\mathcal{U}} x_{ij}$, this problem can be expressed as

$$\text{P2:} \quad \max_{q,x,P,\varepsilon,G} \sum_{i\epsilon\mathcal{B}}\sum_{j\epsilon u} x_{ij} log(c_{ij}) + \sum_{i\epsilon\mathcal{B}} k_i^2 log\left(\sum_{f=1}^{F} p_f q f_i\right) - \sum_{i\epsilon\mathcal{B}} k_i log(k_i) - \eta \sum_{i\epsilon\mathcal{B}} G_i \tag{7}$$

$$\textit{s. t.} \quad C1\ ,\ C2\ ,\ C3\ ,\ C4\ ,\ C5,\ C6,\ C7,\ C8,$$

$$C9: \sum_{j\in u} x_{ij} = k_i, \forall i,$$

in which, $c_{ij} = Blog(1+\gamma_{ij})$.

A. User association and Content Caching

Problem **P2** which is a $non-linear,\ non-convex$ problem, can be defined as a $mixed\ integer\ programing\ problem$ in which a $dual$ subgradient method has been applied to find its optimal solution. Lemma 1 can be applied in order to find the optimal solution for sub-problem P2.1, Considering $\{P, \varepsilon, G\}$, the user association problem is expressed as the following:

$$\textbf{P2.1:} \quad \max_{q,x} \sum_{i\epsilon\mathcal{B}}\sum_{j\epsilon u} x_{ij} log(c_{ij}) + \sum_{i\epsilon\mathcal{B}} k_i^2 log\left(\sum_{f=1}^{F} p_f q f_i\right) - \sum_{i\epsilon\mathcal{B}} k_i log(k_i) \tag{8}$$

$$\textit{s. t.} \quad \text{C1, C2, C4, C5, C6, C9}$$

Lemma 1: if $p_{(1)} \geq \cdots \geq p_{(f)} \geq \cdots \geq p_{(F)}$ indicates the probability degree relevant to the payload (f) requested by a user, the best solution for subproblem P2.1 can be obtained as

$$q_{f_i}^* = \begin{cases} 1, & f_i = (1), \ldots., (L_i) \\ 0, & otherwise \end{cases}, \quad \forall i \in \mathcal{B}. \tag{9}$$

Proof 1: as it is obvious, the goal of solving subproblem P2.1 is maximizing the probability function $\sum_{f=1}^{F} p_f q_f$, regardless of the user association pattern. In accordance with the constraints $C4$ and $C5$, the requested payload can be divided to L_i categories $\mathcal{F}_l (l = 1, \ldots, L_i)$ at base station i, in which the request probability $\mathcal{F}_l$ will be higher compared to $\mathcal{F}_{l+1}$. Hence,

$$\sum_{(f)\in\mathcal{F}_l} q^l_{(f)i} = 1, \sum_{l=1}^{L_i} q^l_{(f)i} = q_{(f)i} \;\; and \;\; \bigcup_{l=L_i} \mathcal{F}_l = \mathcal{F}$$

Here, we have

$$\sum_{f=1}^{F} p_f q_{fi} = \sum_{l=1}^{L_i} \sum_{(f)\in\mathcal{F}l} p_{(f)} q^l_{(f)i} \leq \sum_{l=1}^{L_i} p_{(l)} \left(\sum_{(f)\in\mathcal{F}l} q^l_{(f)i} \right) \Rightarrow \sum_{f=1}^{F} p_f q_{f_i} \leq \sum_{l=1}^{L_i} p_{(l)},$$

Therefore, considering formulation (9), Lemma 1 is proved. Which based on the proved Lemma, we can express problem P2.1 as

$$\widetilde{\mathrm{P}}2.1: \quad \max_{x} \sum_{i\epsilon\mathcal{B}} \sum_{j\epsilon u} x_{ij} log(c_{ij}) + \sum_{i\epsilon\mathcal{B}} k_i^2 log\left(\sum_{f=1}^{L_i} p_{(f)} \right) - \sum_{i\epsilon\mathcal{B}} k_i log(k_i) \tag{10}$$

s. t. C1, C2, C6, C9.

We can also represent problem $\widetilde{\mathrm{P}}2.1$ as a combination of some sub-problems and to find the optimal solution, it's better to work on its dual problem. Based on the Lagrange concepts, the goal function of $\widetilde{\mathrm{P}}2.1$ can be formulated as the following.

$$\begin{aligned} \mathcal{L}(x,k,\mu,\nu) = & \sum_{i\epsilon\mathcal{B}} \sum_{j\epsilon u} x_{ij} log(c_{ij}) + \sum_{i\epsilon\mathcal{B}} k_i^2 log\left(\sum_{f=1}^{L_i} p_{(f)} \right) - \\ & \sum_{i\epsilon\mathcal{B}} k_i log(k_i) - \sum_{j\in u} \mu_j \left(\gamma_{min} - \sum_{i\in\mathcal{B}} x_{ij}\gamma_{uj} \right) - \\ & \sum_{i\in\mathcal{B}} \nu_i \left(\sum_{i\in \mathrm{u}} x_{ij} - k_i \right), \end{aligned} \tag{11}$$

In which, $k = [k_i]$, $\mu = [\mu_j]$, $\upsilon = [\nu_i]$, so that μ_j and ν_i are positive Lagrangian coefficients. So, the dual function $\mathcal{D}(.)$ is defined as (12)

$$\mathcal{D}(\mu,\nu) = \begin{cases} \max\limits_{x,k} \mathcal{L}(x,k,\mu,\nu) \\ s.t. \;\; C2, C6. \end{cases} \tag{12}$$

Hence, we can define the dual problem of $\widetilde{\mathrm{P}}2.1$ (10) as (13)

$$\min_{\mu\geq 0,\nu\geq 0} \mathcal{D}(\mu,\nu). \tag{13}$$

Considering μ_j and ν_i as the dual variables, the optimal value for the goal function will be obviously achieved

$$x^*_{ij} = \begin{cases} 1, & if \quad i = i^* \\ 0, & otherwise \end{cases}, \tag{14}$$

where $i^* = \arg max_i \left(\log(c_{ij}) + \mu_j\gamma_{ij} - \nu_i\right)$. With respect to k_i, the 2^{th} degree derivative of the Lagrange results in

$$\frac{\partial^2 \mathcal{L}}{\partial k_i^2} = 2log\left(\sum_{f=1}^{L_i} p_{(f)} \right) - \frac{1}{k_i}. \tag{15}$$

Because $\sum_{f=1}^{L_i} p_{(f)} \leq 1$, consequently, $\frac{\partial^2 \mathcal{L}}{\partial k_i^2}$ is equal to a negative value. This fact represents that the goal function will be a concave function of k_i. Considering $\frac{\partial^2 \mathcal{L}}{\partial k_i^2}$ equal to zero, the optimal value of k_i can be expressed as $k_i{}^*$

$$k_i^* = -\frac{W\left(-2log\left(\sum_{f=1}^{L_i} p_{(f)}\right) e^{\nu_i - 1}\right)}{2log\left(\sum_{f=1}^{L_i} p_{(f)}\right)}, \tag{16}$$

In this equation, $W(z)$ is the Lambert-W function demonstrating the solution of $z = we^{w}$. According to (14), we can claim that the optimal solution (μ^*, v^*) is not obtainable via differentiable function of $\mathcal{D}(\mu, \nu)$. Hence, we have to apply the sub gradient method as the following

$$\mu_j(t+1) = \left[\mu_j(t) - \delta(t)\left(\sum_{i\in\mathcal{B}} x_{ij}(t)\gamma_{ij} - \gamma_{min}\right)\right]^{+}, \quad (17)$$

$$v_i(t+1) = \left[v_i(t) - \delta(t)\left(k_i(t) - \sum_{j\in u} x_{ij}(t)\right)\right]^{+}, \quad (18)$$

In this formulation, $[a]^{+} = \max\{a, 0\}$, $\delta(t)$ denotes the step size, and t represents the number of iterations. It should be noted that $x_{ij}(t)$ and $k_i(t)$ are updated during each iteration based on equations (14) and (16).

In accordance with the presented analysis, this paper proposes a decentralized cache-enabled user association approach, its functionality has been exhibited through Algorithm 1. The convergence of this algorithm can also be guaranteed because the proposed user association algorithm meets all of the required convergence constraints introduced in [14].

Algorithm 1: User Association with fixed Transmission power

First Step: At the UE domain

1: **if** *t=0,* **then**
2: Initialize $\mu_j(t),\ \forall j$. Each user equipment calculates signal to noise ratio through control channel from all base stations in order to compute c_{ij}.
3: **else**
4: User equipment j takes $v_j(t)$ through base station's broadcast
5: Determination of the serving base station based on
$$i^* = arg\ max_i\ (log(c_{ij}) + \mu_j\gamma_{ij} - v_i)$$
6: update $\mu_j(t)$ based on (17)
7: **end if**
8: $t \rightarrow t+1$
9: Each UE sends the user assiciation demand toward the selected base station

Second Step: At the base station side

1: **if** *t=0,* **then**
2: Initialize $v_j(t),\ \forall j$.
3: **else**
4: Each base station computes $k_j(t)$ based on (15)
5: **Lemma 1** is utilized to estimate the hit probability
6: Each base station takes the updated user association $matrix\ X$.
7: Update $v_j(t)$ based on (18)
8: **end if**
9: $t \rightarrow t+1$
10: Each base station broadcasts $v_j(t)$

B. Resource Allocation

In this section the focus is placed on resource allocation and effective power optimization. After presentation of the proposed user association algorithm, $\{q, x\}$ was obtained with utilizing Algorithm 1. The second main problem, **P2,** is shown as the following:

$$\textbf{P 2.2:} \quad \max_{\mathrm{P},\varepsilon,\mathrm{G}} \sum_{i\epsilon\mathcal{B}}\sum_{j\epsilon u} x_{ij} log(c_{ij}) - \eta \sum_{i\in\mathcal{B}} G_i \quad (19)$$

s. t. *C1, C3, C7, C8.*

Subproblem P2.2 is NP-hard with respect to $\{P_i\}$. We applied a sub gradient method in order to obtain the optimal value for this problem. So, at the first stage, considering ε and G, we try to optimize the power allocation procedure with finding the optimal value for P_i. Here, the sub problem of P2.2-1 is defined as the following.

$$\textbf{P 2.2-1}: \quad \max_{\mathrm{P}} \sum_{i\epsilon\mathcal{B}} \sum_{j\epsilon u} x_{ij} log(c_{ij}) \tag{20}$$

s. t. *C1, C3, C8.*

It should be noted that obtaining the general optimal solution for subproblem $P2.2\text{-}1$ will be very complicated. Hence, a tractable suboptimal framework based on the Newton-Raphson's approach was suggested. It has been proven that the achieved solution is not only efficient, but also quickly converged [15]. As the first stage, with respect to the throughput constraint C_1, the dual function of the main subproblem $P2.2\text{-}1$ is formulated as (21)

$$\widetilde{\mathrm{P}}\ 2.2\text{-}1: \quad \max_{\mathrm{P}} \sum_{i\epsilon\mathcal{B}} \sum_{j\epsilon u} x_{ij} log(c_{ij}) - \sum_{j\in u} \theta_j \left(\gamma_{min} - \sum_{i\in\mathcal{B}} x_{ij}\gamma_{ij} \right) \tag{21}$$

s. t. $0 \le P_i \le \varphi_i, \forall i,$

in which $\{\theta_j\}$ is positive dual coefficient and $\varphi_i = min\{P^i_{max}, G_i + \beta \sum_{i'\in\mathcal{B},i'\neq i} \varepsilon_{i'i} - \sum_{i'\in\mathcal{B},i'\neq i} \varepsilon_{ii'} + E_i\}$ was extracted based on constraints C_3 and C_8. It should be emphasized that the optimal θ_j is achievable utilizing the introduced iterative sub-gradient method according to (17). In continuation, this paper proposes a resource allocation and power optimization solution considering θ_j as a fixed index.

In the problem formulation, $f(P_i)$ illustrates the goal function of subproblem $\widetilde{\mathrm{P}}$ 2.2-1. The 1^{st} and 2^{st} degrees of derivatives of function $f(P_i)$ with respect of P_i are obtained as (22) and (23) respectively.

$$\frac{\partial f(P_i)}{\partial P_i} = \sum_{j\in u} \frac{\gamma_{ij}}{a_{ij}(1+\gamma_{ij})} \frac{x_{ij}}{P_i} - \sum_{i'\in\mathcal{B},i'\neq i} \sum_{j\in u} \frac{h_{ij}\gamma^2_{i'j}}{a_{i'j}(1+\gamma_{i'j})h_{i'j}} \frac{x_{i'j}}{P_{i'}} + \sum_{j\in u} \theta_j \gamma_{ij} \frac{x_{ij}}{P_i} - \sum_{i'\in\mathcal{B},i'\neq i} \sum_{j\in u} \theta_j \frac{h_{ij}\gamma^2_{i'j}x_{i'j}}{h_{i'j}P_{i'}}, \tag{22}$$

and

$$\begin{aligned}\frac{\partial^2 f(P_i)}{\partial P_i^2} = &- \sum_{j\in u} \frac{1+a_{ij}}{a^2_{ij}(1+\gamma_{ij})^2} \left(\frac{\gamma_{ij}}{P_i}\right)^2 x_{ij} \\ &+ \sum_{i'\in\mathcal{B},i'\neq i} \sum_{j\in u} \frac{h^2_{ij}\gamma^3_{i'j}\left(2a_{i'j} + \gamma_{i'j}(a_{i'j}-1)\right)}{h^2_{i'j}P^2_{i'}a^2_{i'j}(1+\gamma_{i'j})^2} x_{i'j} \\ &+ \sum_{i'\in\mathcal{B},i'\neq i} \sum_{j\in u} 2\theta_j \frac{h^2_{ij}\gamma^3_{i'j}}{h^2_{i'j}P^2_{i'}} x_{i'j},\end{aligned} \tag{23}$$

in which $a_{ij} = log(1+\gamma_{ij})$. In order to guarantee achieving the optimal solution, we utilized the customized form of the Newton-Raphson's method as $\Delta P_i = \frac{\partial f(P_i)}{\partial P_i} / \left|\frac{\partial^2 f(P_i)}{\partial P_i^2}\right|$. Therefore, the power optimization formulation will be updated as the following.

$$P_i(\varrho+1) = [P_i(\varrho) + \delta(\varrho)\Delta P_i]_0^{\varphi_i}, \tag{24}$$

In this formulation, $\delta(\varrho)$ represents the size of each stage that can be calculated via backtracking constrained minimization [16]. Index ϱ denotes the iteration number, and the optimal value for the power parameter P_i^* will be achieved after the algorithm's convergence in the last round of iterations. After achieving the solution for subproblem P2.2-1, the pair of (ε, G) will also be renewed via solving the following linear program.

$$\mathbf{P\ 2.2-2}: \quad \min_{\varepsilon, G} \sum_{i \epsilon \mathcal{B}} G_i \tag{25}$$

s. t. C3, C7.

CVX introduced in [17] as a linear programming method for solving iterative problems like *P*2.2−2. Such convex-based tools can also be applied to find the optimal solution for the main problem *P*2.2.

C. Joint Power Allocation and User Association

According to the described framework, an effective joint user association and power optimization approach has been developed to achieve the network's optimal utility in addition to minimize the system's total power consumption, which is summarized in Algorithm 2. Here it should be emphasized that the convergence of the algorithm is guaranteed as long as we have a distinct goal function relevant to both the user association and the power optimization simultaneously, in each iteration [18].

Algorithm 2: Joint User Association & Power Optimization

if $t = 0$
  Initialize $P_i, G_i, E_i, \ \forall i$
else
  Obtain $q_{f_i}$and $x_{ij}(t)$ considering $(\boldsymbol{P}, \boldsymbol{G}, \boldsymbol{\varepsilon})$ applying **algorithm 1**.
  Considering $x_{ij}(t)$ and the relevant pair $(\boldsymbol{G}, \boldsymbol{\varepsilon})$,
  Update the power index $\boldsymbol{P}$ according to the below instructions:
    Loop:
      a) Considering θ_j, loop over $i \in \mathcal{B}$:
        Update P_i based on (24).
        Continue to convergence.
      b) Update θ_j using dual subgradient approach
        Continue to convergence.
  According to the updated power index P, Update G_i and $\mathcal{E}_{ii'}$ through finding the optimal solution for subproblem $\boldsymbol{P2.2-2}$ using convex tools CVX.
  if P2.2-2 converges
    Return the best power optimization strategy
      $(\boldsymbol{q}^*, \boldsymbol{x}^*, \boldsymbol{P}^*, \boldsymbol{\mathcal{E}}^*, \boldsymbol{G}^*)$
    break
  elseif
    $t \rightarrow t + 1$
  end if
end if

IV. Data-Flow Control In Mobile Edge Computing HetNets: Using SNC

In Section III we proposed an energy efficient joint user association & resource allocation algorithm to significantly enhance the system throughput and the total power efficiency of the access network. But, in order to guarantee SLA for content-based services, in the second stage of the paper we suggested an effective content-based approach to improve the end-to-end performance of QoS-aware data-flows in mobile edge computing platforms with random routing.

As exhibited in Figure 3, the agents, that are responsible for connection among the edge clouds, include smart components like edge computing servers. In this model, some of the agents relevant to a distinct route can be applied for multiple streams e.g. agent 1 can be decidated to both the through-flow and interference-flow (a). Although, agents 4 and 7 can be responsible for DPS of through-flow and interference-flow (b). In this scenario we considered 3 types of data-flow each of which is corresponding to a specific QoS-class based on its degree of importance: VoIP $[QoS\text{-}\ class1, (Q1)$, with the highest priority], video stream $[QoS\text{-}\ class2, (Q2)]$, and FTP flow $[QoS\text{-}\ class3, (Q3)]$ with decreasing priority, respectively. We applied a dynamic priority scheduling approach to effectively schedule various QoS-aware data-flows. For example, if we face to a Q1 service (the high priority) demand, when the agent is busy with a Q2 or Q3 data-flow processing, the agent should rapidly suspend the lower priority services and supply the service processing for the Q1 service. It should be noted that services with the identical priority will be processed based on the First-In-First-Out scheduling procedure.

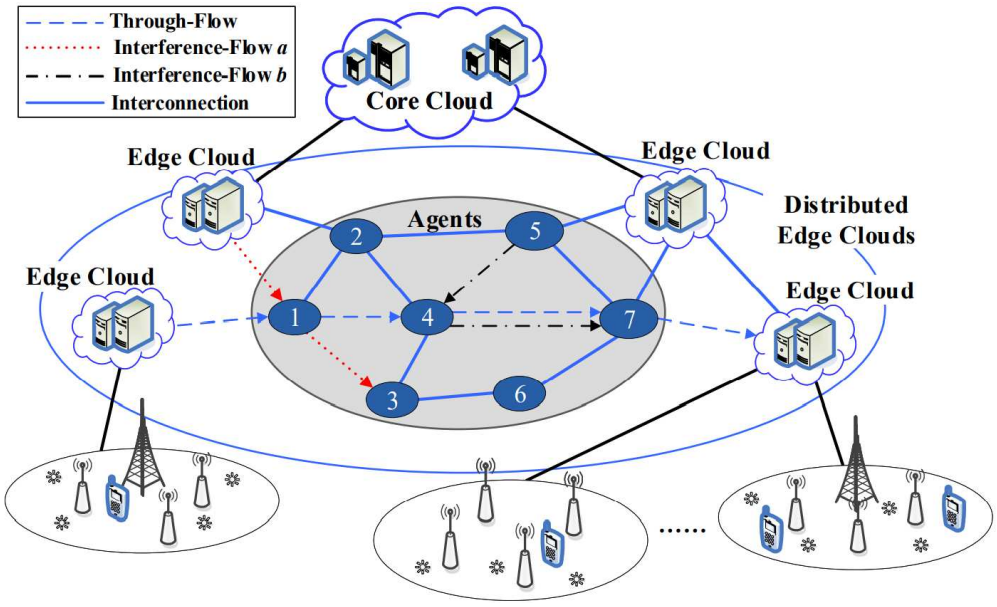

Fig. 3: Configuration of the multi-layer MEC HetNets

The KPI evaluation of mobile edge computing services can be an effective criterion for the network power utility and routing strategies. The current paper proposes an approach for the point-to-point performance analysis of data streams in mobile edge computing networks in accordance with the SNC concepts. Taking into account the random routing in mobile edge computing systems, probability indices were embedded in the proposed assessment framework to consider the randomness characteristics in the evaluation model. Three practical communication scenarios were utilized to analyze the point-to-point functionality of the network data streams. The exclusive priority-scheduling method was employed to analyze these various types of data streams. The data entrance processes of these three streams were taken into account to assess the interference impact on their key performance indicators, service capacity of every node, and data packet transfer model in order to illustrate the correlation between latency, upper bounds of backlogs (UBB), and violation probability (VP) of data streams' QoS indices. According to the analyses and the simulation results, the number of intermediate hops and probability parameters of the interference affected the backlog performance and data stream's latency.

A. System Model

This subsection proposes a hybrid dualhop/multihop scheme for the mobile edge computing network with probability factors and random routing. Figure 4 demonstrates the network data-streaming pattern. Different data streams reach the edge clouds through macro cells or near small cells. The clouds are also interconnected via agents. Moreover, edge clouds are connected to the core cloud.

The data streams of three type of services were submitted by various network entities, including UEs, routers, access points, *etc.* In this scenario, every network entity is known as a network node. The communication path between two nodes might include one hop (singlehop) or multiple hops (multihop). Each node may have several adjacent nodes which can be employed to create a link if they are on the transmission route. The through-flow represents the intended data transmitted through a determined route and interference-flow indicates other shared streams which apply this route. In addition, the services supplied by the network nodes in that route are shared, but it should be noted that the interference flow has a harmful impact on through-flow performance. With the advent of software-defined networking in mobile edge computing, every pair of source and destination nodes can use multiple paths to communicate. Every through-flow selects a random path to the receiver node. In other words, every hop is selected randomly.

In actual scenarios, there are several couples of transmitters/receivers and all of these couples utilize several communication routes which can be selected randomly. The amount of required data for various services of every agent is also random. The interference-flow has no constant effect on the through-flow performance. We considere several major interference scenarios in this paper.

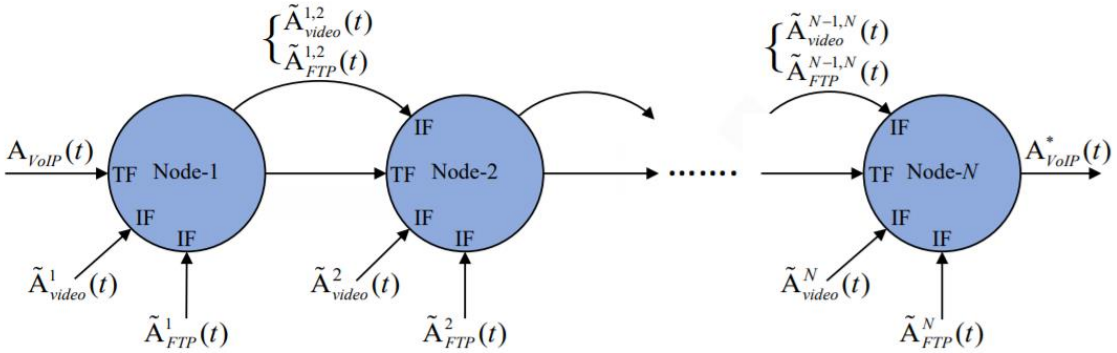

Fig. 4: QoS-aware data flow strategy in the mobile edge computing system

B. Mathematical Modeling of Data-Flows and Service Processes

Based on the proposed framework, stochastic network calculus is employed to develop a mathematical model to analyze the data entrance and service processes in the MEC System. The agent nodes were recognized as service components and the route chosen by the through-flow is configured as service strings in the system's mathematical framework. In this scenario, one of the three data types is regarded as the through-flow, and the other two types are considered as the interference-flow. The performance of every data type is assessed, considering each of them as a through-flow in a specific order. For instance, voice over IP was considered the through-flow in Figure 4, which demonstrates the data flows and network interference.

The moment-generating function was also employed to model the packet transfer state to assess the relevance between end-to-end latency, backlog upper bounds, and violation probability of SLA constraints of different applications in mobile edge computing networks with random routing. These evaluations were done in accordance with the uncomplicated consecutive network topologies demonstrated in Figure 3. As the streaming model illustrated in Figure 4, the method was employed to analyze the through-flow performance is defined as follows:

If the Q1-stream is through-flow, the cumulative traffic arriving into the system during $[s,t]$ is denoted as $A_{Q1}(s,t)$ in which the arrival process of the Q1 stream has Poisson distribution. We indicated the average arrival rate of the Q1-flow as λ_{Q1}. Based on [19], the moment generating function of $A_{Q1}(s,t)$ can be expressed as:

$$M_{A_{Q_1}}(\theta,s,t)=\mathbb{E}\left(e^{\theta A_{Q1}(s,t)}\right)=e^{\lambda_{Q1}(t-s)(e^{\theta}-1)}=\pi_{Q1}^{t-s}(\theta). \quad (26)$$

Now, the Q2-stream is through-flow with the entrance rate (λ_{Q2}). If we consider it fixed, we denote it by r. The cumulative arrival traffic during $[s,t]$ is expressed as $A_{Q2}(s,t)$. Also, the moment generating function of $A_{Q2}(s,t)$ is formulated as below.

$$M_{A_{Q_2}}(\theta,s,t)=\mathbb{E}\left(e^{\theta A_{Q2}(s,t)}\right)=e^{r\theta(t-s)}=\pi_{Q2}^{t-s}(\theta). \quad (27)$$

Because the service type relevant to the Q3 data stream is non-bursty, we can consider Poisson distribution for the flow process.

Subsequently, if the Q3-flow is through-flow, the cumulative arrival traffic during $[s,t]$ will be denoted by $A_{Q3}(s,t)$. Furthermore, the average arrival rate is denoted as λ_{Q3}. The moment generating function of $A_{Q3}(s,t)$ is obtained as (28):

$$M_{A_{Q3}}(\theta,s,t)=\mathbb{E}\left(e^{\theta A_{Q3}(s,t)}\right)=e^{\lambda_{Q3}(t-s)(e^{\theta}-1)}=\pi_{Q3}^{t-s}(\theta). \tag{28}$$

To interference analysis of flow process, the probability parameter set of R-flow on j can be defined as $(P^j_{\mathcal{R}(a)},P^j_{\mathcal{R}(b)},P^j_{\mathcal{R}(c)})$, where $\mathcal{R}$-flow is interference-flow. So, if Q1, Q2, or Q3 flows are interference-flow, their probability parameter sets will be indicated as $(P^j_{Q1},P^j_{Q1},P^j_{Q1})$, $(P^j_{Q2},P^j_{Q2},P^j_{Q2})$, and $(P^j_{Q3},P^j_{Q3},P^j_{Q3})$, respectively.

When Q1-flow is considered as the through-flow, its functionality isn't affected by Q2 and Q3 flows, because it has the topmost priority.

$$O^j_{Q1}(s,t)=0. \tag{29}$$

Theorem 1. If the Q2 data stream is through-flow, the cumulative traffic which interferes with its performance on *j* during $[s,t]$ is obtained as:

$$O^j_{Q1}(s,t)=\sum_{m=1}^{j}P^m_{Q1(a)}P^j_{Q1(c)}\tilde{A}^m_{Q1}(s,t)\prod_{n=m}^{j-1}P^n_{Q1(b)}P^n_{Q1(c)}. \tag{30}$$

Proof: The data-flow Q2 has higher importance than Q3 and less than Q1-flow, so its functionality may only be influenced by Q1 stream. The mathematical framework was proposed to analyse the accumulative volume of Q1 stream which should be considered on *j* during $[s,t]$.

• Case 1: For node 1:

$$I^1_{Q1}(s,t)=P^1_{Q1(a)}P^1_{Q1(c)}\tilde{A}^1_{Q1}(s,t). \tag{31}$$

• Case 2: This case is relevant to the total volume of the Q1-stream on the second node and the stream from the first node is processed at the second node. The cumulative traffic of Q1 stream which should be considered on *j* during $[s,t]$ is achievable by (32).

$$I^2_{Q1}(s,t)=P^1_{Q1(a)}P^1_{Q1(c)}P^1_{Q1(b)}P^2_{Q1(c)}\,\tilde{A}^1_{Q1}\otimes S^j(s,t)+P^2_{Q1(a)}P^2_{Q1(c)}\tilde{A}^2_{Q1}(s,t) \tag{32}$$

It should be noted that the constraint $\lim_{t\to\infty}\frac{S(t)}{t}\geq\lim_{t\to\infty}\frac{A(t)}{t}$ should be met to guarantee the robustness of the network. The total stream of the interference-flow and the service process meet the formula $\alpha\otimes\beta\leq\alpha\wedge\beta$ [20]. Hence, (32) can be expressed as:

$$I^2_{Q1}(s,t)=P^2_{Q1(c)}\left(P^1_{Q1(a)}P^1_{Q1(b)}P^1_{Q1(c)}\,\tilde{A}^1_{Q1}(s,t)+P^2_{Q1(a)}\tilde{A}^2_{Q1}(s,t)\right). \tag{33}$$

• Case 3: On the third node, the total data stream of Q1is equal to:

$$\begin{aligned}I^3_{Q1}(s,t)=&P^2_{Q1(b)}P^2_{Q1(c)}P^3_{Q1(c)}\\&*\left(P^1_{Q1(a)}P^1_{Q1(b)}P^1_{Q1(c)}\tilde{A}^1_{Q1}(s,t)\right.\\&\left.+P^2_{Q1(a)}\tilde{A}^2_{Q1}(s,t)\right)\\&+P^3_{Q1(a)}P^3_{Q1(c)}\tilde{A}^3_{Q1}(s,t).\end{aligned} \tag{34}$$

• Case *j*: considering the presented mathematical analysis, the total data stream of Q1-flow on *j* can be obtained as

$$I^j_{Q1}(s,t)=\sum_{m=1}^{j}P^m_{Q1(a)}P^j_{Q1(c)}\tilde{A}^m_{Q1}(s,t)\prod_{n=m}^{j-1}P^n_{Q1(b)}P^n_{Q1(c)}, \tag{35}$$

Hence,

$$O^j_{Q2}(s,t)=\sum_{m=1}^{j}P^m_{Q1(a)}P^j_{Q1(c)}\tilde{A}^m_{Q1}(s,t)\prod_{n=m}^{j-1}P^n_{Q1(b)}P^n_{Q1(c)}.$$

Theorem 2. If we consider the Q3 stream as the through-flow, the cumulative traffic on node *j*, is obtained via (36).

$$O_{Q3}^{j}(s,t) = \sum_{m=1}^{j} P_{Q1(a)}^{m} P_{Q1(c)}^{j} \tilde{A}_{Q1}^{m}(s,t) \prod_{n=m}^{j-1} P_{Q1(b)}^{n} P_{Q1(c)}^{n} + \sum_{m=1}^{j} P_{Q2(a)}^{m} P_{Q2(c)}^{j} \tilde{A}_{Q2}^{m}(s,t) \prod_{n=m}^{j-1} P_{Q2(b)}^{n} P_{Q2(c)}^{n}. \quad (36)$$

Proof: When the Q3 data stream is through-flow, the Q1 and Q2 data streams concurrently play an interference role for the quality of the Q3-flow. In order to study the interference effect of Q1-flow on the functionality of Q2-flow, the cumulative traffic of Q1 data stream should be processed on relay j during $[s,t]$ which is shown as (35). Likewise, the cumulative traffic of Q2 data stream required to be processed on j during $[s,t]$ can be calculated as:

$$I_{Q2}^{j}(s,t) = \sum_{m=1}^{j} P_{Q2(a)}^{m} P_{Q2(c)}^{j} \tilde{A}_{Q2}^{m}(s,t) \prod_{n=m}^{j-1} P_{Q2(b)}^{n} P_{Q2(c)}^{n}. \quad (37)$$

Hence, based on the superposition property, the cumulative traffic that affects the utility of Q3-flow will be achievable as the following.

$$O_{Q3}^{j}(s,t) = I_{Q1}^{j}(s,t) + I_{Q2}^{j}(s,t) = \sum_{m=1}^{j} P_{Q1(a)}^{m} P_{Q1(c)}^{j} \tilde{A}_{Q1}^{m}(s,t) \prod_{n=m}^{j-1} P_{Q1(b)}^{n} P_{Q1(c)}^{n} + \sum_{m=1}^{j} P_{Q2(a)}^{m} P_{Q2(c)}^{j} \tilde{A}_{Q2}^{m}(s,t) \prod_{n=m}^{j-1} P_{Q2(b)}^{n} P_{Q2(c)}^{n}. \quad (38)$$

The cumulative service provided by j during [s, t] is demonstrated by $S^{j}(s,t)$. Note that the service data rate of all agents is considered to be the same for all three types of streams, which is denoted by a fixed value 'C'. Based on [21], the service process of node j will be as the following.

$$S^{j}(s,t) = C(t-s), \quad (39)$$

Therefore, the moment generating function of the service process is achievable as (40):

$$\bar{M}_{S^j}(\theta,s,t) = \mathbb{E}\left(e^{-\theta S^{j}(s,t)}\right) = e^{-\theta(C(t-s))} = \varphi^{t-s}(\theta). \quad (40)$$

Now, the service process of the three QoS-class data streams can be analyzed. Each data-flow is constant as through-flow.

• Q1-flow: When the Q1 data stream is through-flow with the best priority, other data-flows cannot affect this service. The service process of this data flow on relay j can be obtained by the following:

$$S_{Q1}^{j}(s,t) = S^{j}(s,t), \quad (41)$$

And its moment generating function is shown as:

$$\bar{M}_{S_{Q2}^{j}}(\theta,s,t) = \bar{M}_{S^j}(\theta,s,t) = \varphi^{t-s}(\theta). \quad (42)$$

• Q2-flow: suppose that the Q2-flow is the through-flow. In this condition, Q3 data flow cannot affect this service but it can be affected by Q1 data stream. Based on this rule, the service process of Q2 data stream on node j is obtained as (43):

$$S_{Q2}^{j}(s,t) \geq S^{j}(s,t) - O_{Q2}^{j}(s,t). \quad (43)$$

If Q1-flow is an interference-flow, the entrance rate of Q1 data stream at node j can be calculated as $\tilde{\lambda}_{Q1}^{j}$. Based on (27), the MGF of $\tilde{A}_{Q1}^{j}(s,t)$ is equal to:

$$M_{\tilde{A}_{Q1}^{j}}(\theta,s,t) = e^{\tilde{\lambda}_{Q1}^{j}(t-s)(e^{\theta}-1)} = \tilde{\pi}_{Q1(j)}^{t-s}(\theta). \quad (44)$$

$$M_{O_{Q2}^{j}}(\theta,s,t) = \prod_{m=1}^{j} e^{\tilde{\lambda}_{Q2}^{j}(t-s)\left(e^{P_{Q1(a)}^{m} P_{Q1(c)}^{j} \prod_{n=1}^{j-1} \theta P_{Q1(b)}^{n} P_{Q1(c)}^{n} - 1}\right)} \quad (45)$$

$$= \left(\prod_{m=1}^{j} e^{\tilde{\lambda}_{Q1}^{j}\left(e^{P_{Q1(a)}^{m} P_{Q1(c)}^{j} \prod_{n=1}^{j-1} \theta P_{Q1(b)}^{n} P_{Q1(c)}^{n}} - 1\right)} \right)^{(t-s)} = \rho_j^{t-s}(\theta).$$

Based on (29) and (44), we obtained the moment generating function of $O_{Q2}^{j}(s,t)$, Hence, the moment generating function of (42) can be converted to the following.

$$\bar{M}_{S_{Q2}^{j}}(\theta,s,t) \le \bar{M}_{S^j - O_{Q2}^{j}}(\theta,s,t) = \bar{M}_{S^j}(\theta,s,t) \,.\, \bar{M}_{O_{Q2}^{j}}(\theta,s,t)) = \varphi^{t-s}(\theta)\rho_j^{t-s}(\theta). \quad (46)$$

Likewise, the Q3 data stream with the worst priority is affected by the Q1 and Q2 data streams. If the Q3-flow is the through-flow, the service process will be formulated as the following:

$$S_{Q3}^{j}(s,t) \ge S^{j}(s,t) - O_{Q3}^{j}(s,t). \quad (47)$$

If we consider Q2-flow as interference-flow, the arrival rate $\tilde{\lambda}_{Q2}^{j}$ at node j is fixed and it can be indicated by $\tilde{r}^{j}$. Based on (28), we can extract the Moment Generating Function of $\tilde{A}_{Q2}^{j}(s,t)$ as (48):

$$M_{\tilde{A}_{Q2}^{j}}(\theta,s,t) = e^{\theta \tilde{r}^{j}(t-s)} = \tilde{\pi}_{Q2(j)}^{t-s}(\theta) \,. \quad (48)$$

$$M_{I_{Q2}^{j}}(\theta,s,t) = \prod_{m=1}^{j} e^{\theta \tilde{r}^{m}(t-s) P_{Q2(a)}^{m} P_{Q2(c)}^{j} \prod_{n=1}^{j-1} P_{Q2(b)}^{n} P_{Q2(c)}^{n}} \quad (49)$$

$$= \left(\prod_{m=1}^{j} e^{\tilde{r}^{j} P_{Q2(a)}^{m} P_{Q2(c)}^{j} \prod_{n=1}^{j-1} P_{Q2(b)}^{n} P_{Q2(c)}^{n}} \right)^{(t-s)\theta} = \psi_j^{t-s}(\theta).$$

According to (37) and (49), the Moment Generating Function of $I_{Q2}^{j}(s,t)$ is achievable. So, the moment generating function of $O_{Q3}^{j}(s,t)$ is expressed as:

$$M_{O_{Q3}^{j}}(\theta,s,t) = M_{I_{Q1}^{j}}(\theta,s,t) \,.\, M_{I_{Q2}^{j}}(\theta,s,t) \;= \rho_j^{t-s}(\theta)\psi_j^{t-s}(\theta) \,. \quad (50)$$

Also, based on (47) and (50), the moment generating function of $S_{Q3}^{j}(s,t)$ is obtained as:

$$\bar{M}_{S_{Q3}^{j}}(\theta,s,t) \le \bar{M}_{S^j}(\theta,s,t) \,.\, \bar{M}_{O_{Q3}^{j}}(\theta,s,t) = \varphi^{t-s}(\theta)\rho_j^{t-s}(\theta)\psi_j^{t-s}(\theta). \quad (51)$$

C. Data-Flow Control Performance: Network KPI

To evaluate the data-flow performance of this model, we considered some key performance indicators such as delay upper bound (DUB) and backlog upper bounds (BUB) for all of the defined data streams with different QoS requirements. In order to exhibit the relationship between these KPIs, we derived the closed-form formulation of services in accordance with their arrivals and processes.

Theorem 3. If we consider $\Re$ as the through-flow, the total traffic of this data stream during [0, t] is indicated as $A_{\Re}(t)$, and the overall service provided by the system is denoted by $S_{\Re}^{net}(t)$. Furthermore, the total traffic of the $\Re$ stream going out from the network during [0, t] is equal to $A_{\Re}^{*}(t)$. So, If we define N as the number of network entities on the service chain, the relationship between the DUB and the VP meets the following formulation.

$$\mathbb{P}\{D_{\Re}(t) > x\} \le M_{A_{\Re} \oslash S_{\Re}^{net}}(\theta, t+x, t). \quad (52)$$

Proof:

$$\begin{aligned}
&\mathbb{P}\{D_{\Re}(t) > x\} = \mathbb{P}\{A_{\Re}(t) > A_{\Re}^{*}(t+x)\} \\
&= \mathbb{P}\{A_{\Re}(t) - A_{\Re}^{*}(t+x) > 0\} \\
&\le \mathbb{P}\{A_{\Re}(t) - A_{\Re} \otimes S_{\Re}^{net}(t+x) > 0\} \\
&= \mathbb{P}\left\{A_{\Re}(t) - \inf_{0 \le s \le t+x} \{A_{\Re}(s) + S_{\Re}^{net}(s, t+x)\} > 0\right\} \\
&= \mathbb{P}\left\{\sup_{0 \le s \le t+x} \{A_{\Re}(t) - A_{\Re}(s) - S_{\Re}^{net}(s, t+x)\} > 0\right\} \\
&= \mathbb{P}\{A_{\Re} \oslash S_{\Re}^{net}(t+x, t) > 0\} \le M_{A_{\Re} \oslash S_{\Re}^{net}}(\theta, t+x, t).
\end{aligned}$$

In accordance with [22], the moment generating function can be converted to deconvolution form as

$$M_{A_{\Re}\emptyset S_{\Re}^{net}}(\theta, t+x, t) \le \frac{\left(\pi_{\Re}(\theta)\right)^{-x}}{\left(1-\emptyset_{\Re}(\theta)\pi_{\Re}(\theta)\right)^{N}}, \quad (53)$$

Where $\emptyset_{\Re}(\theta) = \underset{1 \le j \le N}{max}\left\{\left(\bar{M}_{S_{\Re}^{j}}(\theta, s, t)\right)^{\frac{1}{t-s}}\right\}$

Hence, the relationship between the DUB and the VP of the three data streams with different QoS-class requirements are expressed as follows:

• Q1-flow: The Q1-data stream has the best priority compared to the other two data-flows.

$$\mathbb{P}\{D_{Q1}(t) > x\} \le \frac{e^{-\lambda_{Q1}x(e^{\theta}-1)}}{\left(1-\emptyset_{Q1}(\theta)e^{\lambda_{Q1}(e^{\theta}-1)}\right)^{N}}, \quad (54)$$

$$where\ \emptyset_{Q1}(\theta) = \underset{1 \le j \le N}{max\{\varphi(\theta)\}}.$$

• Q2-flow: The priority of Q2-data stream is more than Q3-flow and lower than Q1-flow.

$$\mathbb{P}\{D_{Q2}(t) > x\} \le \frac{e^{\theta(-rx)}}{\left(1-\emptyset_{Q2}(\theta)e^{\theta r}\right)^{N}}, \quad (55)$$

$$where\ \emptyset_{Q1}(\theta) = \underset{1 \le j \le N}{max\{\varphi(\theta)p_j(\theta)\}}.$$

• Q3-flow: as mentioned before, the priority of Q3-data stream is lower than two other data flows.

$$\mathbb{P}\{D_{Q3}(t) > x\} \le \frac{e^{-\lambda_{Q3}x(e^{\theta}-1)}}{\left(1-\emptyset_{Q3}(\theta)e^{\lambda_{Q3}(e^{\theta}-1)}\right)^{N}}, \quad (56)$$

$$where\ \emptyset_{Q3}(\theta) = \underset{1 \le j \le N}{max\{\varphi(\theta)\rho_j(\theta)\psi_j(\theta)\}}.$$

Theorem 4. If we consider $\Re$ -flow as the through-flow, the relationship between the BUB and VP of the $\Re$ -flow results in:

$$\mathbb{P}\{B_{\Re}(t) > x\} \le e^{-\theta x} M_{A_{\Re}\emptyset S_{\Re}^{net}}(\theta, t, t). \quad (57)$$

$$Proof: B_{\Re}(t) = A_{\Re}(t) - A_{\Re}^{*}(t)$$

$$= A_{\Re}(t) - A_{\Re} \otimes S_{\Re}^{net}(t) = A_{\Re}(t) - \underset{0 \le s \le t}{inf\{A_{\Re}(S) + S_{\Re}^{net}(s,t)\}}$$

$$= \underset{0 \le s \le t}{sup\{A_{\Re}(s,t) - S_{\Re}^{net}(s,t)\}} = A_{\Re}\emptyset S_{\Re}^{net}(t,t).$$

Based on the Chernoff theory, we can conclude that the relationship between the BUB and VP is equal to:

$$\mathbb{P}\{B_{\Re}(t) > x\} \le e^{-\theta x} M_{B_{\Re}}(\theta) = e^{-\theta x} M_{A_{\Re}\emptyset S_{\Re}^{net}}(\theta, t, t). \quad (58)$$

Subsequently, relationship between the BUB and VP of the three data streams is formulated as below.

• Q1-flow

$$\mathbb{P}\{B_{Q1}(t) > x\} \le \frac{e^{-\theta x}}{\left(1-\emptyset_{Q1}(\theta)e^{\lambda_{Q1}(e^{\theta}-1)}\right)^{N}} \quad (59)$$

• Q2-flow

$$\mathbb{P}\{B_{Q2}(t) > x\} \le \frac{e^{-\theta x}}{\left(1-\emptyset_{Q2}(\theta)e^{\theta r}\right)^{N}}. \quad (60)$$

• Q3-flow

$$\mathbb{P}\{B_{Q3}(t) > x\} \leq \frac{e^{-\theta x}}{\left(1 - \emptyset_{Q3}(\theta) e^{\lambda_{Q3}(e^{\theta}-1)}\right)^{N}} . \quad (61)$$

V. RESULTS

In this section the simulation scenarios are presented and the results are interpreted to evaluate the performance of the proposed algorithms and to exhibit the effectiveness of different optimization approaches. In this regard, some of the main key performance indicators like total energy efficiency, network utilization, network data rate and fairness index are assessed under different scenarios.

A. Simulation Scenarios

The simulation scenarios were implemented in Python so that the optimal solution can be achieved by applying CPLEX as a powerful flexible optimizer for high-performance mathematical programming. After obtaining the optimal solution, the deep-learning outputs (decision variables) determine UE association, small cell activity pattern (which small cells are required to turn on or off), and the backhaul routing mechanism for every UE demand, as well as other deployment items. In the simulation scenarios, the macro base station was placed in the center of the macro cell, the radius of which was set at 300 m, and the radius of every small cell was 15 m. The total network bandwidth was considered 40 MHz, so the operational frequency was set to 3.5 GHz, and the system included 30 independent carriers. The small cells were also random distributed within the macro cell range.

Table 1 illustrates the simulation characteristics in which all of the parameters are based on 3GPP standard introduced in [23]. The user throughput threshold ($R_{Threshold}$) was defined considering $P^i(0) = P_P^{max}$, whereas $b^i(0) = 0$ and $\beta = 1$ were considered for ease of use. Although $R_{Threshold}$ can be relevant to the SC positions and user distribution, it is applied only to test the sensitivity of the proposed algorithm to the UE throughput limitations. The threshold can also be adjusted based on the circumstances.

To verify the proposed *user association* and *power optimization* algorithms in terms of EE, the simulations were done based on a HetNet layout comprising numerous small cells, one macro base station, and 150 pieces of user equipment in addition to several backhaul links (for connecting baseband unit (BBU) and the core network) in a line-of-sight communications. The length of every backhaul link ranges between 50 m and 350 m. The small cells were scattered in a 10×10 grid with a 250-m inter-site distance. In this model, small cells were able to apply multiple-orthogonal carriers to avoid cross-tier interference, although this presumption cannot be very precise considering the limited number of existing carriers in a dense environment with numerous small cells. The frequency-reuse technique will be analyzed in future works, in which the cross-tier interference among small cells can be calculated based on the on/off status of the cells. The user equipment are distributed based on two distinct distribution patterns within the macro cell coverage area, 1) uniform (U) and 2) hotpot (Hs) [24], in which 70% of the user equipment are concentrated within a 100-m radius surrounding the two random small cells. In addition, regardless of the distribution pattern, 20 random drops are assumed for each scenario, without any severe blockage probability. The two user distribution patterns were deployed as:

- Hotspot (Hs) user distribution: The user equipment are distributed randomly in a circular area within a 10–30 m radius of a small cell and a 30–50 m radius of the macro base station.
- Random (Rn) user distribution: The user equipment are scattered randomly within the macro cell coverage area.

In this paper, the channel gain included line-of-sight pathloss, log-normal shadowing, and fast Rayleigh fading. The propagation model is crosswave which is compatible with a dense urban profile. The Rayleigh fading model is appropriate for cellular radio propagation due to the existence of many reflection points of multipath connections. The channel gain is considered as an independent variable with unit-mean and exponential distribution. The required users' data rate ranged between 2–25 MB/s, whereas the bandwidth of every millimeter-wave backhaul link is 3.5 GHz. Other variables

were initialized as: $Rx\ loss = Tx\ loss = 5\ dB$;
$G_{TX} = G_{RX} = [V - band: 44\ dBi, E - band: 46\ dBi]$; $L_{margin} = 12\ dB$ and $NF = [V - band: 5\ dB, E - band: 6.5\ dB]$. Table 1 shows the rest of the simulation variables. The outcomes show the average results of less than 1000 independent simulations. The performance indices used in these analyses include the total energy efficiency, network utilization, system data rate and fairness index. In addition, all of the assessment indexes have been analyzed under the maximum-transmission-power conditions. The proposed approach called "Energy Efficient Dynamic Flow control for User association & Power optimization (*EEFUP*)" was compared with some other energy efficient flow control and resource optimization algorithms which we briefly introduce them as the following:

Table 1. Main simulation parameters

Parameter	Value
HetNet Configuration	Hexagonal network, 3-sectored BSs
Small cell distribution pattern	uniform (U) and hotspot (Hs),
Operational Frequency (GHz)	3.5 Gigahertz
BW (MHz)	2*20 Megahertz
Power backoff	2.8 dB
Maximum transmission power of macro-BS	45 dBm
Codec configuration	Adaptive multi-rate
Inter site spacing	250 m
Hopping method	Synthesized frequency hopping (RF)
$\boldsymbol{Rx\ loss}$ & $\boldsymbol{Tx\ loss}$	5 dB
Propagation model	Cross-wave propagation model
Scheduler	Fair
Maximum allowed iteration	1000
$\boldsymbol{L_{margin}}$	12 dBm
Operational FREQ of mmWave BH link	3.5 GHz
The learning factor $\boldsymbol{c_1 = c_2}$	1.54
Maximum weighting factor $\boldsymbol{\omega_{max}}$	0.85
Minimum weighting factor $\boldsymbol{\omega_{min}}$	0.45
Power consumption pattern of macro-BS $\boldsymbol{P_m^o}$	110 w
Power consumption pattern of small-BS $\boldsymbol{P_s^o}$	5.5 w
Small cell radius	15 m
Antenna Type	APE4517R0-0698X_CO-P45_03T

Random *Allocation (RA)*- which is a discrete resource allocation scheme without any optimality in which the resource allocation is done merely based on the received demands. *Fixed Power Allocation (FPA)*- in this scheme, all base stations equally share their maximum transmission power to all the associated user equipment and backhaul links and subsequently, each user equipment and backhaul link applies all the allocated resources. *User Grouping and Fractional Transmit Power Control (UG-FTPC)*- This algorithm acts upon user grouping and fractional MTP control. Based on the UG-FTPC scheme, the UEs are categorized as d_v groups in accordance with their channel gains, and each UE can only share subchannels with the UEs which are not in the same group with it. In this power allocation scheme, more power can be assigned to the UEs with poor channel quality to guarantee fairness consideration [25]. *Successive Codebook Ordering Assignment (SCOA)*- to better evaluate the performance of our proposed algorithms, the functionality of *EEFUP* was also compared to SCOA. The hybrid power domain sparse code NOMA (*PD-SCMA*) that applies both power domain NOMA (*PD-NOMA*) and sparse code MA (*SCMA*) for an uplink hierarchical heterogeneous network. The functionality of SCOA is based on channel quality ordering criterion includes opportunistic MUE-SUE pairing (*OMSP*) for UE pairing (UP), and a QoS-aware resource allocation (*QARA*) for resource management (*RM*) [26].

B. General Descriptions of Figures

Figure 5 illustrates the network's power efficiency performance with respect to the number of small cells by estimating the error variance at 0.05 for the channel gain with 5 to 25 users associated to each small cell and 10 to 50 small cells per macro cell. The upper bound of power for every small

base station is equal to 23 dB/mWatt. Accordingly, the system's energy efficiency increases with the number of small cells and UEs. The system's energy efficiency is clearly 2.5 and 2 times greater for 50 small cells than the energy efficiency for 20 and 30 small cells, respectively. In other words, the larger the number of base stations and UEs are, the better the performance of the proposed algorithm will be. In this scenario, the power efficiency curve EEFUP has a steeper slope, and the power efficiency improvement is more evident. According to this figure, EEFUP outperformed other optimization algorithms, because SOCA and UG-FTPC failed to use completely all of the accessible cooperative resources.

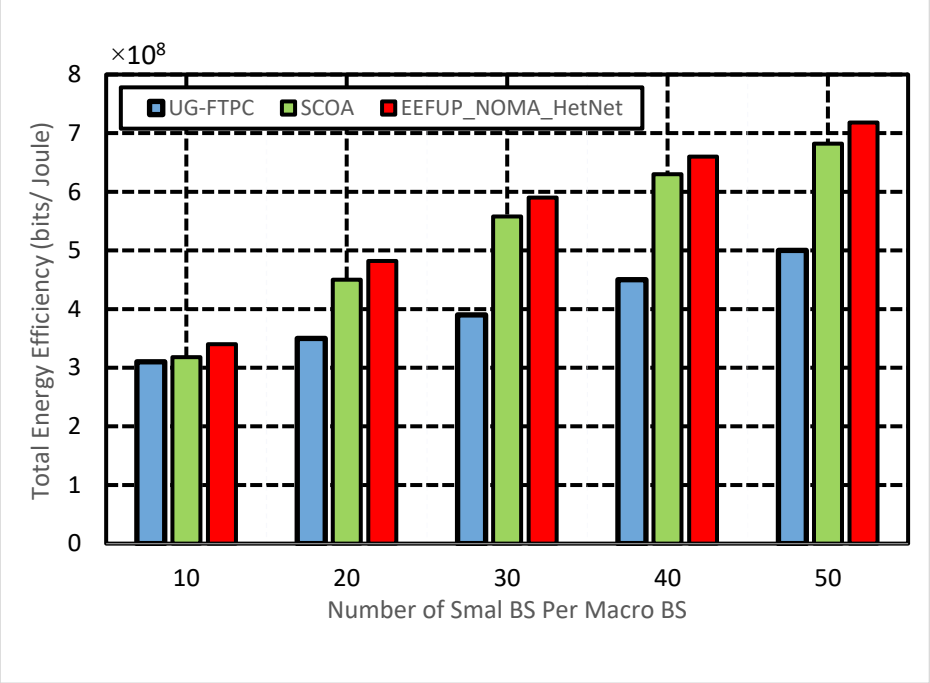

Fig. 5. Energy efficiency *vs.* number of small cells.

Figure 6 illustrates the energy efficiency performance considering different limitations of the maximum available power of small cells in the optimization algorithms. The proposed energy efficient user association & power optimization algorithm, by jointly optimizing transmission power and dynamic flow control is shown as EEFUP. In this scenario, the proposed algorithm was compared with three other algorithms, the first of which was SCOA algorithm which optimizes the transmission power and cell association bias to maximize total energy efficiency. The second algorithm is UG-FTPC in which the UEs are categorized as d_v groups in accordance with their channel gains, and each UE can only share subchannels with the UEs which are not in the same group with it. This algorithm only optimizes the transmission power of small cells without considering throughput requirements. The Random Power Allocation as the third algorithm makes no change in transmission energy or cell association bias. According to the figure, EEFUP algorithm outperformed the other algorithms in terms of energy efficiency. Furthermore, when the maximum transmission power limitation is below 20 db/mmWatt, there is an upward trend in the EEFUP. In other words, users receive more energy with respect to the maximum transmission power limitation within this range; therefore, more users are handovered from macro cells to small cells, which leads to more spectrum resources and improves the network's sum rate. At the same time, power can be saved by changing the power transmission and turning off some small cells. It is evident that increasing the transmission power threshold doesn't have any positive effect on the total energy efficiency performance due to increased imposed interference to the network.

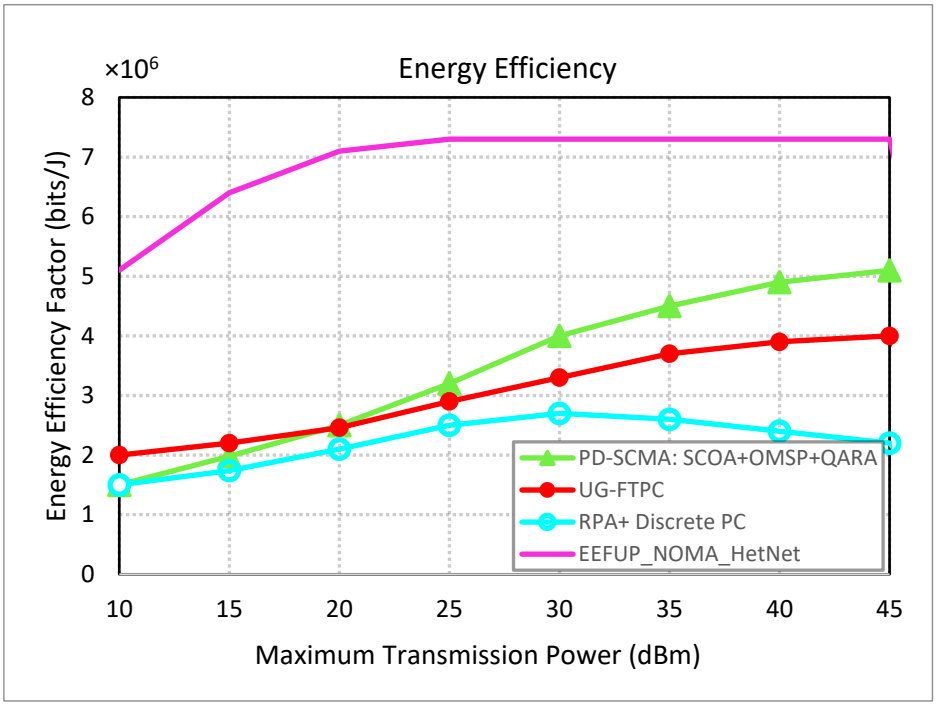

Fig. 6. The performance of energy efficiency (EE) under different solutions

Figure 7 represents the throughput evaluation for different algorithms. According to the simulation results, the throughput of EEFUP increased until the transmission power threshold reached 25 db/mmWatt. It then remained constant, because the simultaneous optimization of carrier matching and transmission power kept the network interference low. In addition, more macro users can be allocated to small cells by optimizing the cell association bias; therefore, they will have wider spectra resources to utilize. Nevertheless, increasing the maximum transmission power (MTP) can intensify intercellular interference; as a result, the throughput remains constant.

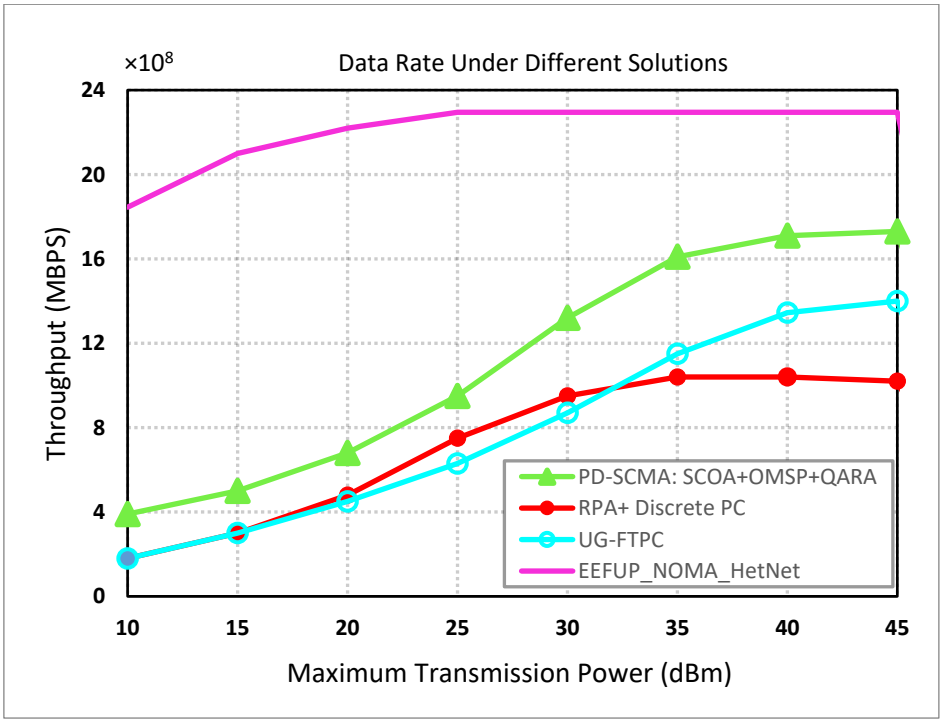

Fig. 7. The performance of throughput under different solutions

According to Figures 6 and 7, if the range of MTP is below 25 db/mmWatt, the throughput gain will be higher than the power consumption gain. This causes improvement of the energy efficiency performance. Nevertheless, when the upper bound of the transmission power exceeds 25 db/mmWatt, the throughput gain does not increase in parallel with the power consumption; therefore, the energy efficiency performance remains constant.

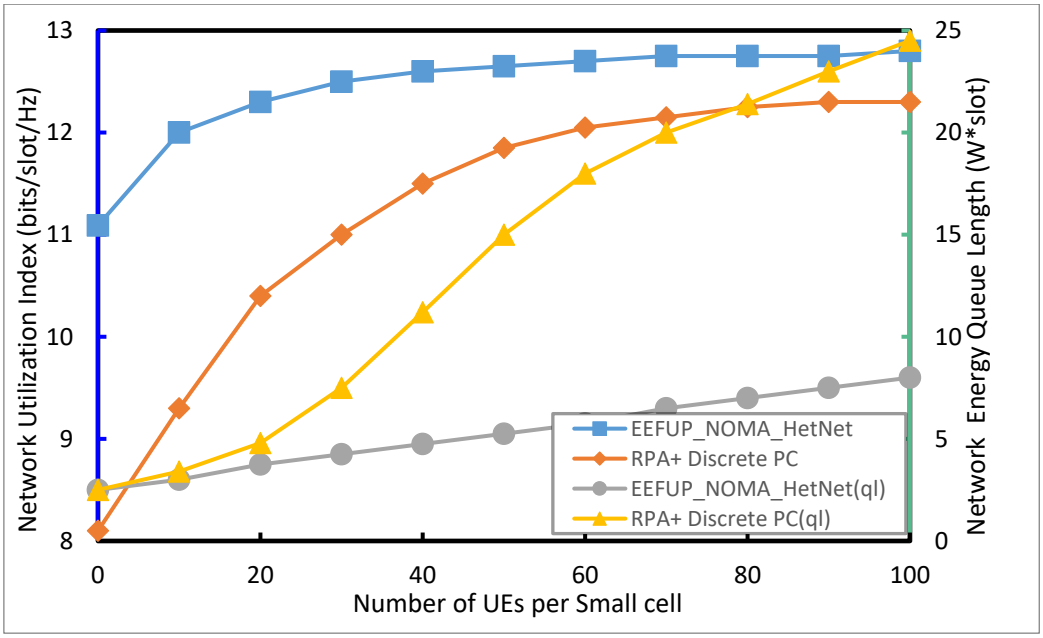

Fig. 8. Average network utility and energy queue length versus UE density

Fig. 8 exhibits the network utilization and energy queue length based on the number of UEs per small cell. In this scenario $\beta = 0.85$ and UE density is considered variable from 1 to 100 per small cell. As it is obvious, the average network utilization and the energy queue length increase with increasing the number of UEs. In this simulation, the average network utilization index of EEFUP is rapidly converged to its optimal value. Considering the same UE density, not only the average network utilization index of EEFUP is higher than RPA but also its average energy queue length is significantly less than RPA. This is relevant to the energy cooperation and effective power utilization capabilities of EEFUP compared to the Random Power Allocation.

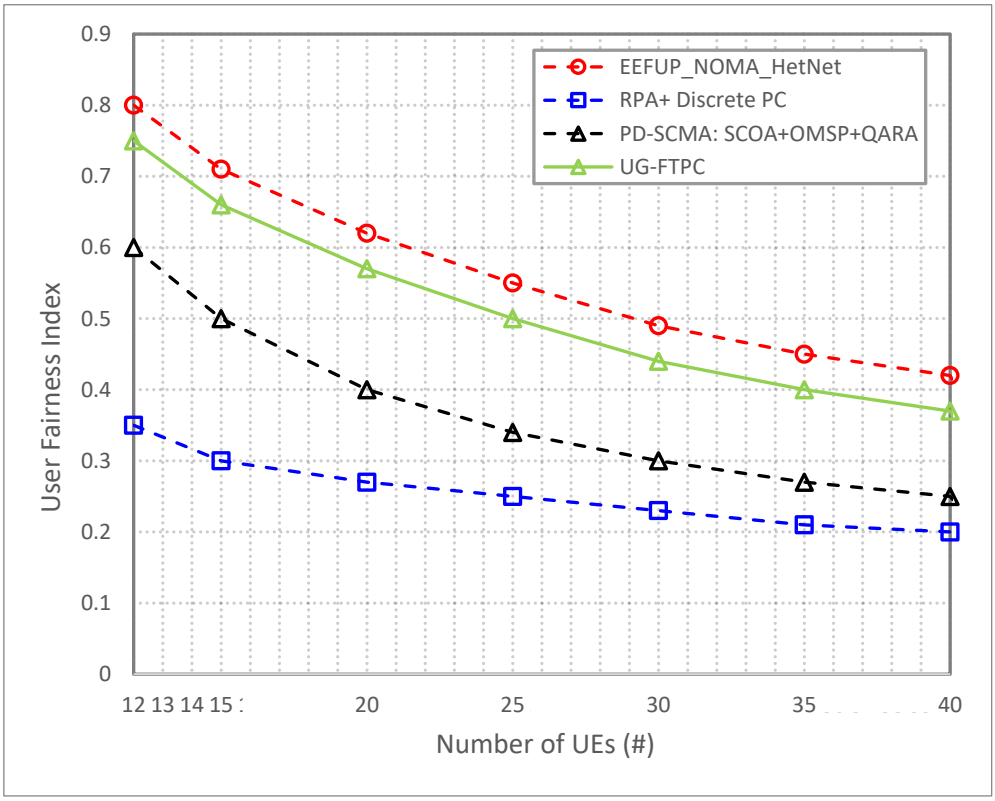

Fig. 9. User fairness index vs number of the users

Fig. 9 exhibits the simulation results and draws a comparison in terms of the fairness. The evaluation is done based on the UE fairness index versus the UE density. In this scenario we considered Jain fairness method [27] in which the user throughput should be monitored within 60 various time slots. According to the achieved results, EEFUP has higher fairness than UG-FTPC, which itself has higher fairness than the *Successive Codebook Ordering Assignment* (SCOA) method. Therefore, EEFUP can provide higher throughput for UEs located in poor coverage areas than the values obtained from UG-FTPC and SCOA methods. The reason is that the proposed approach with high levels of fairness in the system will probably allocate a considerable part of resources to the UEs whose channel status is worse than others. Accordingly, the UG-FTPC algorithm was fairer than SCOA, because most of the transmission energy is allocated to the cell-edge users whose channel statuses are worse than others. Based on UG-FTPC scheme, the UEs are categorized to d_v groups in accordance with their channel gains, and each UE can only share subchannels with the UEs which are not in the same group with it. In this power allocation scheme, more power can be assigned to the UEs with poor channel quality to guarantee fairness consideration.

ABBREVIATIONS

MEC: Mobile edge computing; SNC: stochastic network calculus; QoS: Quality of service; UE: User equipment; SDN: software-defined networking; MGF: moment-generating function; NGMN: Next-generation mobile networks; NOMA: Non-orthogonal multipe access; MBS: Macro base station; HetNet: Heterogeneous network; UA: User association; VoIP: Voice over IP; KPI: Key performance indicator; DUB: delay upper bound; BUB: backlog upper bounds; MTP: Maximum transmission power; RPA: Random power allocation.

DECLARATIONS

Funding

No specific funding has been provided for this paper.

Conflicts of interest:

The author of the paper declares that he doesn't have any conflict of interest.

Availability of data and material
Data sharing not applicable to this article as no datasets were generated or analysed during the current study
Code availability
Not applicable for this paper
Animal Research
Not applicable to this paper
Clinical Trials Registration
Not applicable to this paper
Author Contribution
This paper has one author which have done all of the preparation stages relevant to the paper (writing, simulations, methods, literature,…)
Consent to Participate and Publish
The athor declares his consent to participate and publish this research in the journal
Plant Reproducibility
Not applicable to this paper

VI. DISCUSSION

In this paper power optimization in cache-enabled cooperative heterogeneous networks was investigated in which the joint user association and power control algorithms were presented to maximize the total network throughput and the fairness index while keeping the grid power utilization low. The proposed approach concurrently optimizes fronthaul and backhaul operations for mobile edge computing networks using deep-learning in the Core and the Access network layers. In the access layer, the energy of macro and small cells with various cache sizes is provided through conventional grid networks and renewable power resources in which the power can be shared among base stations via grid networks. The paper presents an energy-efficient stochastic network calculus (SNC) framework to control MEC network data flows. In accordance with the data entrance processes with different QoS-classes, closed-form problem was formulated to determine the correlation between resource utilization and the violation probability of data flows through random routing. This paper also analyzed the fairness and violation probability of the three stream types as well as the effect of hops density and the sets of interference-flow probabilities on the QoS-aware through-flow performance. The simulation results exhibit that the proposed approach can effectively improve the user throughput and total power efficiency while guaranteeing the acceptable fairness level for uniform and hotspot UE distributions. It also proved that the energy utilization index and the system data rate can be significantly improved. Moreover, some other important aspects were introduced for the future works to develop this model. e.g. online heuristics can be applied to achieve the estimated solutions. Also, the traffic model is expected to contain the variations considering time and delay constraints.

REFERENCES

[1] Zhou, Huan, Hui Wang, Xiuhua Li, and Victor CM Leung. "A survey on mobile data offloading technologies." IEEE Access 6 (2018): 5101-5111.

[2] Ning, Zhaolong, Kaiyuan Zhang, Xiaojie Wang, Lei Guo, Xiping Hu, Jun Huang, Bin Hu, and Ricky YK Kwok. "Intelligent edge computing in internet of vehicles: a joint computation offloading and caching solution." IEEE Transactions on Intelligent Transportation Systems (2020).

[3] Yu, Shuai, Rami Langar, Xiaoming Fu, Li Wang, and Zhu Han. "Computation offloading with data caching enhancement for mobile edge computing." IEEE Transactions on Vehicular Technology 67, no. 11 (2018): 11098-11112.

[4] Liu, Ling, Yiqing Zhou, Jinhong Yuan, Weihua Zhuang, and Ying Wang. "Economically optimal MS association for multimedia content delivery in cache-enabled heterogeneous cloud radio access networks." IEEE Journal on Selected Areas in Communications 37, no. 7 (2019): 1584-1593.

[5] Wang, Zenan, Jiao Zhang, and Tao Huang. "Determining Delay Bounds for A Chain of Virtual Network Functions Using Network Calculus." IEEE Communications Letters (2021).

[6] Smith, J. MacGregor. "Open Queueing Network Algorithms $$ f {\bigl (G (V, E)\bigr)} $$." In Introduction to Queueing Networks, pp. 181-259. Springer, Cham, 2018.

[7] Ma, Shengcheng, Xin Chen, Zhuo Li, and Ying Chen. "Performance Evaluation of URLLC in 5G Based on Stochastic Network Calculus." Mobile Networks and Applications (2019): 1-13.

[8] Asshad, Muhammad, Adnan Kavak, Kerem Küçük, and Sajjad Ahmad Khan. "Using moment generating function for performance analysis in non-regenerative cooperative relay networks with max-min relay selection." AEU-International Journal of Electronics and Communications 116 (2020): 153066.

[9] Qian, Yuwen, Shi Li, Long Shi, Jun Li, Feng Shu, Dushantha Nalin K. Jayakody, and Jinhong Yuan. "Cache-enabled MIMO power line communications with precoding design in smart grid." IEEE Transactions on Green Communications and Networking 4, no. 1 (2019): 315-325.

[10] Liu, Jinbo, and Shaohui Sun. "Energy efficiency analysis of cache-enabled cooperative dense small cell networks." IET Communications 11, no. 4 (2017): 477-482.

[11] Gu, Shushi, Yan Tan, Ning Zhang, and Qinyu Zhang. "Energy-Efficient Content Placement with Coded Transmission in Cache-Enabled Hierarchical Industrial IoT Networks." IEEE Transactions on Industrial Informatics (2020).

[12] Wang, Yitu, Wei Wang, Ying Cui, Kang G. Shin, and Zhaoyang Zhang. "Distributed packet forwarding and caching based on stochastic network utility maximization." IEEE/ACM Transactions on Networking 26, no. 3 (2018): 1264-1277.

[13] Chien, Su Fong, Charilaos C. Zarakovitis, Qiang Ni, and Pei Xiao. "Stochastic asymmetric blotto game approach for wireless resource allocation strategies." IEEE Transactions on Wireless Communications 18, no. 12 (2019): 5511-5528.

[14] Zhang, Haijun, Site Huang, Chunxiao Jiang, Keping Long, Victor CM Leung, and H. Vincent Poor. "Energy efficient user association and power allocation in millimeter-wave-based ultra dense networks with energy harvesting base stations." IEEE Journal on Selected Areas in Communications 35, no. 9 (2017): 1936-1947.

[15] Muhammed, Alemu Jorgi, Zheng Ma, Panagiotis D. Diamantoulakis, Li Li, and George K. Karagiannidis. "Energy-efficient resource allocation in multicarrier NOMA systems with fairness." IEEE Transactions on Communications 67, no. 12 (2019): 8639-8654.

[16] Chen, Xihan, Yunlong Cai, Qingjiang Shi, Minjian Zhao, Benoit Champagne, and Lajos Hanzo. "Efficient resource allocation for relay-assisted computation offloading in mobile-edge computing." IEEE Internet of Things Journal 7, no. 3 (2019): 2452-2468.

[17] Yang, Gang, Xinyue Xu, and Ying-Chang Liang. "Resource allocation in NOMA-enhanced backscatter communication networks for wireless powered IoT." IEEE Wireless Communications Letters 9, no. 1 (2019): 117-120.

[18] Cui, Gaofeng, Xiaoyao Li, Lexi Xu, and Weidong Wang. "Latency and Energy Optimization for MEC Enhanced SAT-IoT Networks." IEEE Access 8 (2020): 55915-55926.

[19] Huang, Haojun, Wang Miao, Geyong Min, Jialin Tian, and Atif Alamri. "NFV and Blockchain Enabled 5G for Ultra-Reliable and Low-Latency Communications in Industry: Architecture and Performance Evaluation." IEEE Transactions on Industrial Informatics (2020).

[20] Ma, Zhongyu, Jie Cao, Qun Guo, Xiangwei Li, and Hongfeng Ma. "QoS-oriented joint optimization of concurrent scheduling and power control in millimeter wave mesh backhaul network." Journal of Network and Computer Applications 174 (2021): 102891.

[21] Lu, Zhonghai, and Xueqian Zhao. "xMAS-based QoS analysis methodology." IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems 37, no. 2 (2017): 364-377.

[22] Yang, Guang, Ming Xiao, and Zhibo Pang. "Delay analysis of traffic dispersion with Nakagami-m fading in millimeter-wave bands." In 2018 IEEE Wireless Communications and Networking Conference (WCNC), pp. 1-6. IEEE, 2018.

[23] Jao, Chin-Kuo, Chun-Yen Wang, Ting-Yu Yeh, Chun-Chia Tsai, Li-Chung Lo, Jen-Hsien Chen, Wei-Chen Pao, and Wern-Ho Sheen. "WiSE: a system-level simulator for 5G mobile networks." IEEE Wireless Communications 25, no. 2 (2018): 4-7.

[24] Zhang, Haijun, Haisen Zhang, Keping Long, and George K. Karagiannidis. "Deep Learning Based Radio Resource Management in NOMA Networks: User Association, Subchannel and Power Allocation." IEEE Transactions on Network Science and Engineering 7, no. 4 (2020): 2406-2415.

[25] Huang, Xiaoge, Dongyu Zhang, She Tang, Qianbin Chen, and Jie Zhang. "Fairness-based distributed resource allocation in two-tier heterogeneous networks." IEEE Access 7 (2019): 40000-40012.

[26] Chege, Simon, and Tom Walingo. "Energy efficient resource allocation for uplink hybrid power domain sparse code nonorthogonal multiple access heterogeneous networks with statistical channel estimation." Transactions on Emerging Telecommunications Technologies 32, no. 1 (2021): e4185.

[27] Pratap, Ajay, Rajiv Misra, and Sajal K. Das. "Maximizing fairness for resource allocation in heterogeneous 5g networks." IEEE Transactions on Mobile Computing (2019).