Efficient Resource Scheduling and Dispatch of Mobile Cell Sites to Improve 5G Performance

You-Chiun Wang
Department of Computer Science and Engineering
National Sun Yat-sen University
Kaohsiung, Taiwan
ycwang@cse.nsysu.edu.tw

Ching-Ting Chu
Department of Computer Science and Engineering
National Sun Yat-sen University
Kaohsiung, Taiwan
m083040032@student.nsysu.edu.tw

***Abstract*—To fulfil the rapid growth of traffic demands and diverse network applications, 5G base stations (called gNBs) are being widely deployed. Due to some activities like sport events, the service area would be crowded with many *user equipments (UEs)*. Thus, gNBs may not have enough spectrum resources to serve UEs, leading to low throughput and high packet loss. This paper uses a few *mobile cell sites (MCSs)* to solve the problem, which are gNBs equipped on vehicles and can flexibly move to serve UEs. We propose an *efficient resource scheduling and MCS dispatch (ESMD)* scheme, which helps each gNB allot resources to UEs to raise throughput. Then, we pick candidate locations for MCSs to move to help share loads of busy gNBs. Specifically, we adopt the Pareto-optimal concept to find one-to-one assignments between MCSs and these locations. Through simulations, we show that the ESMD scheme efficiently improves network performance in terms of throughput, packet loss, and the outage ratio.**

***Keywords*—5G, dispatch, mobile cell site, Pareto optimality, resource scheduling.**

I. Introduction

Many people nowadays rely on mobile networks for communications and Internet access. They usually use network applications that are bandwidth-consuming and delay-sensitive, like video streaming and on-line meeting. Thus, 3GPP defines the global specifications for the fifth generation (5G) mobile networks. 5G features include massive network capacity, high transmission speeds, very low latency, better user experience, more reliability, and increased availability [1]. By 2023, there will be 13.1 billion *user equipments (UEs)* in the global, and 1.4 billion of these UEs will be 5G capable [2].

3GPP defines two deployment modes for 5G networks: *non-standalone (NSA)* and *standalone (SA)*. The NSA mode allows 5G base stations (called *gNBs*) to remain reliant on the 4G core network for management and signaling, and 4G base stations continue to operate. This mode provides a transitional platform for operators and users alike. On the other hand, the SA mode pairs gNBs with a cloud-native 5G core network, so it enables all features of 5G. In this paper, we consider the SA mode.

In a service area, the density and traffic demands of UEs may not necessarily be stable [3]. For example, there are many UEs in office areas or campuses in workdays, but they become almost empty in holidays. Besides, there will be an influx of UEs to some hotspot regions due to special activities such as singing concerts and sports events. In this case, the gNBs may not have enough spectrum resources to serve all UEs, thereby lowering their throughput and increasing the outage ratio.

One promising solution is to use *mobile cell sites (MCSs)*, which are gNBs equipped on vehicles (e.g., trucks or UAVs). They can flexibly move to serve UEs, thereby sharing workloads of busy gNBs. This paper proposes an *efficient resource scheduling and MCS dispatch (ESMD)* scheme to improve the performance of a 5G network by using MCSs. Each gNB allocates resources to UEs with the aim of maximizing throughput. Then, we check if some UEs do not get enough resources to satisfy their demands, and dispatch MCSs to serve them. To do so, we find candidate locations for MCSs, and calculate a maximum matching between MCSs and these locations such that the matching is Pareto-optimal. Simulation results show that ESMD can substantially raise network throughput, reduce packet loss, and lower the outage ratio.

II. Related Work

In the literature, there are different MCS issues discussed. The work [4] maximizes the service time of each MCS by reducing its traveling distance. Becvar et al. [5] show that using MCSs can be an efficient substitute to ultra-dense small cell deployment, especially when users move in crowds. Shinbo et al. [6] employ MCSs for temporary mobile communications in a disaster area. Given a set of ground terminals, the study [7] uses the minimum MCSs to cover them, where each terminal is within the communication range of at least one MCS. The work [8] finds the optimal moving direction for each MCS to achieve the highest spectral efficiency. Huang et al. [9] model the MCS placement problem as a sparse optimization problem, which is solved by the reweighted l_1-norm algorithm. Ree et al. [10] propose a decentralized key management scheme to provide security in a mobile network with MCSs. As can be seen, the above studies have different objectives with ours.

Several studies also dispatch MCSs to serve UEs. Without a signal-strength radio map, the work [11] finds the MCS's target location based on fine-grained line-of-sight information to raise throughput. Alzenad et al. [12] select visiting locations for an MCS, such that it can use the minimum power to serve the maximum UEs. The study [13] takes account of the power, bandwidth, and position of an MCS to maximize throughput. However, these studies aim at a single MCS. The work [14] dispatches multiple MCSs to share the loads of gNBs based on a *genetic algorithm (GA)*. As discussed later in Section V, our ESMD scheme outperforms the GA-based method in terms of throughput, packet loss, and the outage ratio.

III. System Model

We consider a service area seamlessly covered by 5G cells in the SA mode. Let $\hat{\mathcal{B}}$ be the set of their gNBs. In each cell, a gNB takes charge of allocating the spectrum resource to UEs, whose smallest unit is one *resource block (RB)*. Depending on the channel bandwidth, the gNB provides a different number of RBs in one *transmission time interval (TTI)*. Every RB can be allotted to only one UE, and its capacity is decided by the UE's channel quality [15].

UEs are randomly distributed in the service area, but parts of them may congregate in some small regions, namely *hotspots*. A set $\hat{\mathcal{M}}$ of MCSs will move in the area and offer extra RBs to UEs, where $|\hat{\mathcal{M}}| \ll |\hat{\mathcal{B}}|$. Given the position and traffic demand of each UE, our objective is to help each gNB in $\hat{\mathcal{B}}$ efficiently allot RBs to its UEs (i.e., resource scheduling) and adaptively move MCSs in $\hat{\mathcal{M}}$ to share loads of gNBs (i.e., MCS dispatch), such that the network performance is maximized. Specifically, we adopt throughput, packet loss, and the outage ratio of UEs as the metrics for performance evaluation.

IV. The Proposed ESMD Scheme

Our ESMD scheme consists of two modules. In the *scheduling module*, each gNB in $\hat{\mathcal{B}}$ first allots RBs to UEs in its cell. If a UE cannot get enough RBs to fulfill its demand, we call it a *non-satisfied-yet (NSY)* UE. Based on the distribution of NSY UEs, we find a set $\hat{\mathcal{L}}$ of candidate locations in the service area. Afterward, the *dispatching module* constructs a bipartite graph to reveal the relationship between the MCSs in $\hat{\mathcal{M}}$ and the locations in $\hat{\mathcal{L}}$. Then, a maximum Pareto-optimal matching is found in the bipartite graph for assigning MCSs to move to the selected locations, so as to serve NSY UEs and share the loads of gNBs. Below, we detail each module.

A. The Scheduling Module

Each UE u_i estimates the *signal-to-interference-plus-noise-ratio (SINR)* from each gNB $b_j \in \hat{\mathcal{B}}_i$ as follows [16]:

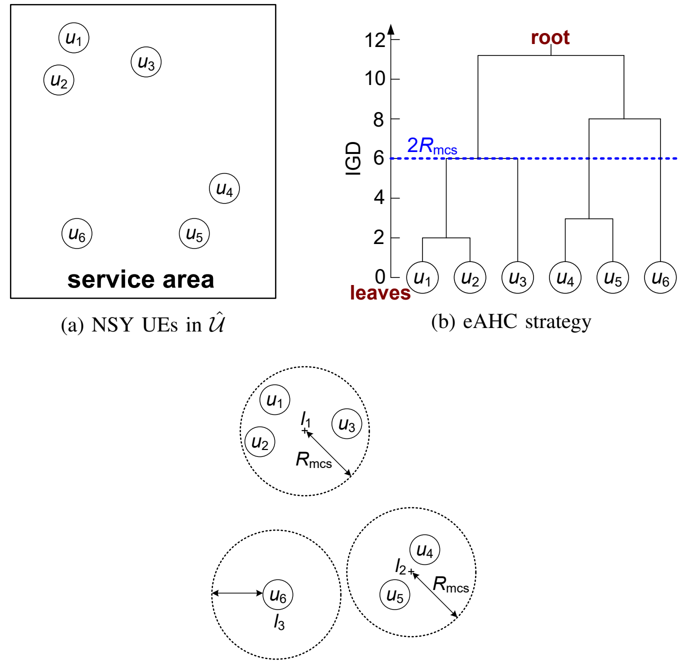
$$\mathrm{SINR}(u_i, b_j) = \frac{\tilde{P}_{j,i}}{\sigma + \sum_{b_k \in \hat{\mathcal{B}}_i, b_k \neq b_j} \tilde{P}_{k,i}}, \quad (1)$$

where $\hat{\mathcal{B}}_i$ is the subset of gNBs in $\hat{\mathcal{B}}$ whose signals are captured by u_i, $\tilde{P}_{j,i}$ is the amount of b_j's power received by u_i, and σ is the environmental noise. The SINR is converted to a *channel quality indicator (CQI)* [17]. Then, u_i sorts each gNB in $\hat{\mathcal{B}}_i$ based on the CQI decreasingly, and selects the first gNB, say, b_j to ask for service. If b_j has no RBs, it is removed from $\hat{\mathcal{B}}_i$, and then u_i selects the next gNB. Otherwise, b_j will accept u_i's request. The above procedure is repeated until either u_i finds a gNB for service or $\hat{\mathcal{B}}_i$ becomes empty. In the latter case, u_i is marked as one NSY UE.

When a gNB b_j accepts u_i's request, b_j chooses a *modulation and coding (M&C)* scheme to send u_i's data according to its CQI. Let d_i be u_i's traffic demand (in bits/s) and ζ_i be the maximum number of data bits carried by one RB (depending on u_i's associated M&C scheme). Then, the number of RBs that b_j should allocate to u_i to fulfill its demand is [18]:

$$r_i^{\min} = \arg\min_{r_i} \{\zeta_i r_i \geq d_i t / 1000\}, \quad (2)$$

where t is the length of a TTI (measured in ms). However, if b_j has fewer than $r_i^{\min}$ available RBs, b_j allocates the residual RBs to u_i. In this case, u_i is also treated as an NSY UE.



(a) NSY UEs in $\hat{\mathcal{U}}$ (b) eAHC strategy

(c) candidate locations in $\hat{\mathcal{L}}$

Fig. 1. Example of grouping NSY UEs and finding candidate locations.

After allocating RBs, we find a set $\hat{\mathcal{L}}$ of candidate locations for MCSs to move to serve NSY UEs. Let $\hat{\mathcal{U}}$ be the set of all NSY UEs. We adopt the *enhanced agglomerative hierarchical clustering (eAHC)* strategy in [19] to divide $\hat{\mathcal{U}}$ into groups, so that each group of NSY UEs can be covered by one MCS. In particular, each UE in $\hat{\mathcal{U}}$ is initially added to a different group. Afterward, we iteratively merge two groups $\hat{\mathcal{G}}_\alpha$ and $\hat{\mathcal{G}}_\beta$ such that they have the shortest *inter-group distance (IGD)*, which is the distance between the two farthest UEs $u_i \in \hat{\mathcal{G}}_\alpha$ and $u_j \in \hat{\mathcal{G}}_\beta$. The iteration is repeated until the shortest IGD overtakes $2R_{\text{mcs}}$, where R_{mcs} is the radius of an MCS's communication range. Fig. 1 presents an example, where $R_{\text{mcs}} = 3$. Fig. 1(a) gives the distribution of NSY UEs in $\hat{\mathcal{U}}$. Then, Fig. 1(b) shows that eAHC divides $\hat{\mathcal{U}}$ into three groups: $\hat{\mathcal{G}}_1 = \{u_1, u_2, u_3\}$, $\hat{\mathcal{G}}_2 = \{u_4, u_5\}$, and $\hat{\mathcal{G}}_3 = \{u_6\}$.

In each group $\hat{\mathcal{G}}_k$, the best location to place an MCS is the *geometric center (GC)* of all NSY UEs. More concretely, let (x_i, y_i) be the coordinate of an NSY UE u_i. Then, the GC's coordinate can be calculated as follows:

$$(x_i^{\mathbf{gc}}, y_i^{\mathbf{gc}}) = \left(\frac{1}{|\hat{\mathcal{G}}_k|} \sum\nolimits_{u_i \in \hat{\mathcal{G}}_k} x_i, \frac{1}{|\hat{\mathcal{G}}_k|} \sum\nolimits_{u_i \in \hat{\mathcal{G}}_k} y_i \right). \quad (3)$$

Then, we can add $(x_i^{\mathbf{gc}}, y_i^{\mathbf{gc}})$ to $\hat{\mathcal{L}}$. Fig. 1(c) shows an example, where l_1, l_2, and l_3 are the GCs of $\hat{\mathcal{G}}_1$, $\hat{\mathcal{G}}_2$, and $\hat{\mathcal{G}}_3$, respectively.

B. The Dispatching Module

Given both $\hat{\mathcal{M}}$ and $\hat{\mathcal{L}}$, we build a bipartite graph $(\hat{\mathcal{V}}, \hat{\mathcal{E}}) = (\hat{\mathcal{M}} \cup \hat{\mathcal{L}}, \hat{\mathcal{M}} \times \hat{\mathcal{L}})$, where all MCSs in $\hat{\mathcal{M}}$ and all locations in $\hat{\mathcal{L}}$ are converted into vertices in $\hat{\mathcal{V}}$. Edges in $\hat{\mathcal{E}}$ connect vertices between $\hat{\mathcal{M}}$ and $\hat{\mathcal{L}}$. Then, the MCS dispatching problem can be translated into the problem of finding a matching $\mathcal{H}$ such that (1) $\mathcal{H}$ contains the maximum pairs of MCSs and locations and (2) $\mathcal{H}$ is Pareto-optimal. Here, objective (1) is to fully utilize MCSs to serve UEs and objective (2) indicates that one cannot find another matching whose result is better than that of $\mathcal{H}$.

For objective (1), we use the Hopcroft-Karp algorithm [20] to find a maximum matching from the graph. More concretely, a vertex $v_i \in \hat{\mathcal{V}}$ is *free* if $(v_i, v_j) \notin \mathcal{H}$ for every $(v_i, v_j) \in \hat{\mathcal{E}}$. Besides, a path $\mathcal{P} = \{(v_1, v_2), (v_2, v_3), \cdots, (v_{i-1}, v_i)\}$ is called an *augmenting path* if both v_1 and v_i are free, and $\mathcal{P}$'s edges alternatively appear in $\hat{\mathcal{E}} - \mathcal{H}$ and $\mathcal{H}$. Then, a maximum matching $\mathcal{H}$ can be found by the following steps:

1. At the beginning, $\mathcal{H}$ contains an (arbitrary) edge in $\hat{\mathcal{E}}$.
2. Find augmenting path $\mathcal{P}$ for $\mathcal{H}$. We generate a matching $\mathcal{H}' = \mathcal{H} \oplus \mathcal{P}$, which contains the symmetric difference of $\mathcal{H}$'s edges and $\mathcal{P}$'s edges. Then, we replace $\mathcal{H}$ by $\mathcal{H}'$.
3. Repeat step 2 until no augmenting path can be found.

Theorem 1. *Let $|\hat{\mathcal{M}}| = \xi_M$ and $|\hat{\mathcal{L}}| = \xi_L$. Finding a maximum matching $\mathcal{H}$ takes $O(\xi_M \times \xi_L(\sqrt{\xi_M + \xi_L}))$ time.*

Proof: The Hopcroft-Karp algorithm spends $O(|\hat{\mathcal{E}}| \cdot |\hat{\mathcal{V}}|^{\frac{1}{2}})$ time. Since the bipartite graph is complete, we can derive that $|\hat{\mathcal{E}}| = |\hat{\mathcal{M}} \times \hat{\mathcal{L}}| = |\hat{\mathcal{M}}| \times |\hat{\mathcal{L}}| = \xi_M \times \xi_L$ and $|\hat{\mathcal{V}}| = |\hat{\mathcal{M}} \cup \hat{\mathcal{L}}| = |\hat{\mathcal{M}}| + |\hat{\mathcal{L}}| = \xi_M + \xi_L$. Thus, the theorem is verified. ■

For objective (2), we change some pairs in $\mathcal{H}$ to make it Pareto-optimal. Let $\tilde{N}(m_i, l_j)$ be the number of NSY UEs that an MCS $m_i \in \hat{\mathcal{M}}$ can serve when it moves to a location $l_j \in \hat{\mathcal{L}}$. Then, we can define the *preference* of m_i on matchings. In particular, Let m_i be paired with locations l_{j_1} and l_{j_2} in two matchings $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively, where $l_{j_1}, l_{j_2} \in \hat{\mathcal{L}}$. Then, m_i prefers $\mathcal{H}_1$ to $\mathcal{H}_2$ if either condition is met:

- $\tilde{N}(m_i, l_{j_1}) > \tilde{N}(m_i, l_{j_2})$, which means that m_i can serve more NSY UEs at location l_{j_1}.
- $\tilde{N}(m_i, l_{j_1})$ is equal to $\tilde{N}(m_i, l_{j_2})$. Besides, if m_i moves to location l_{j_1}, the CQI of no UE degrades, and the CQI of at least one UE can rise. In other words, the channel quality of some UEs will improve. However, this situation does not happen when m_i moves to location l_{j_2}.

For convenience, we denote by "$f(m_i, l_{j_1}) > f(m_i, l_{j_2})$" if any of the two conditions holds.

Definition 1. *Let us denote by "$\mathcal{H}_1 >_{\mathrm{p}} \mathcal{H}_2$" if no MCS prefers $\mathcal{H}_2$ to $\mathcal{H}_1$, and some MCSs prefer $\mathcal{H}_1$ to $\mathcal{H}_2$. Then, a matching $\mathcal{H}$ is said to be Pareto-optimal if and only if we cannot find another matching, say, $\mathcal{H}'$ such that $\mathcal{H}' >_{\mathrm{p}} \mathcal{H}$.*

Then, we perform two operations to change some pairs in $\mathcal{H}$ to make it become Pareto-optimal.

TABLE I
PARAMETERS OF GNBS AND MCSS.

parameter	gNB	MCS
number	100	3, 5, 7
cell range	400 m	200 m
transmitted power	46 dBm	30 dBm
channel bandwidth	100 MHz	80 MHz
RBs offered per TTI	273	217

Trade-in-free checking operation: Suppose that an MCS m_i is paired with a location l_j in $\mathcal{H}$. If there exists an unpaired location l_k such that $f(m_i, l_k) > f(m_i, l_j)$, we replace pair (m_i, l_j) with pair (m_i, l_k) in $\mathcal{H}$. This operation is repeated until no such replacement can be performed.

Coalition-free checking operation: Suppose that $\mathcal{H}$ contains a sequence of pairs (m_{i_1}, l_{j_1}), (m_{i_2}, l_{j_2}), $\cdots$, (m_{i_k}, l_{j_k}) such that $f(m_{i_1}, l_{j_2}) > f(m_{i_1}, l_{j_1})$, $f(m_{i_2}, l_{j_3}) > f(m_{i_2}, l_{j_2})$, $\cdots$, $f(m_{i_{k-1}}, l_{j_k}) > f(m_{i_{k-1}}, l_{j_{k-1}})$, and $f(m_{i_k}, l_{j_1}) > f(m_{i_k}, l_{j_k})$. Then, we remove pairs (m_{i_1}, l_{j_1}), (m_{i_2}, l_{j_2}), $\cdots$, (m_{i_k}, l_{i_k}) from $\mathcal{H}$ and add pairs (m_{i_1}, l_{j_2}), (m_{i_2}, l_{j_3}), $\cdots$, $(m_{i_{k-1}}, l_{j_k})$, and (m_{i_k}, l_{j_1}) to $\mathcal{H}$. This operation is repeated until no such sequence can be found.

The study [21] proves that by performing trade-in-free and coalition-free checking operations on a matching $\mathcal{H}$, $\mathcal{H}$ must be Pareto-optimal. Then, for each pair (m_i, l_j) in $\mathcal{H}$, we can dispatch MCS m_i to move to location l_j for serving NSY UEs.

Theorem 2. *It requires $O(\xi_M(\xi_L \log_2 \xi_L))$ time to run trade-in-free and coalition-free checking operations on matching $\mathcal{H}$.*

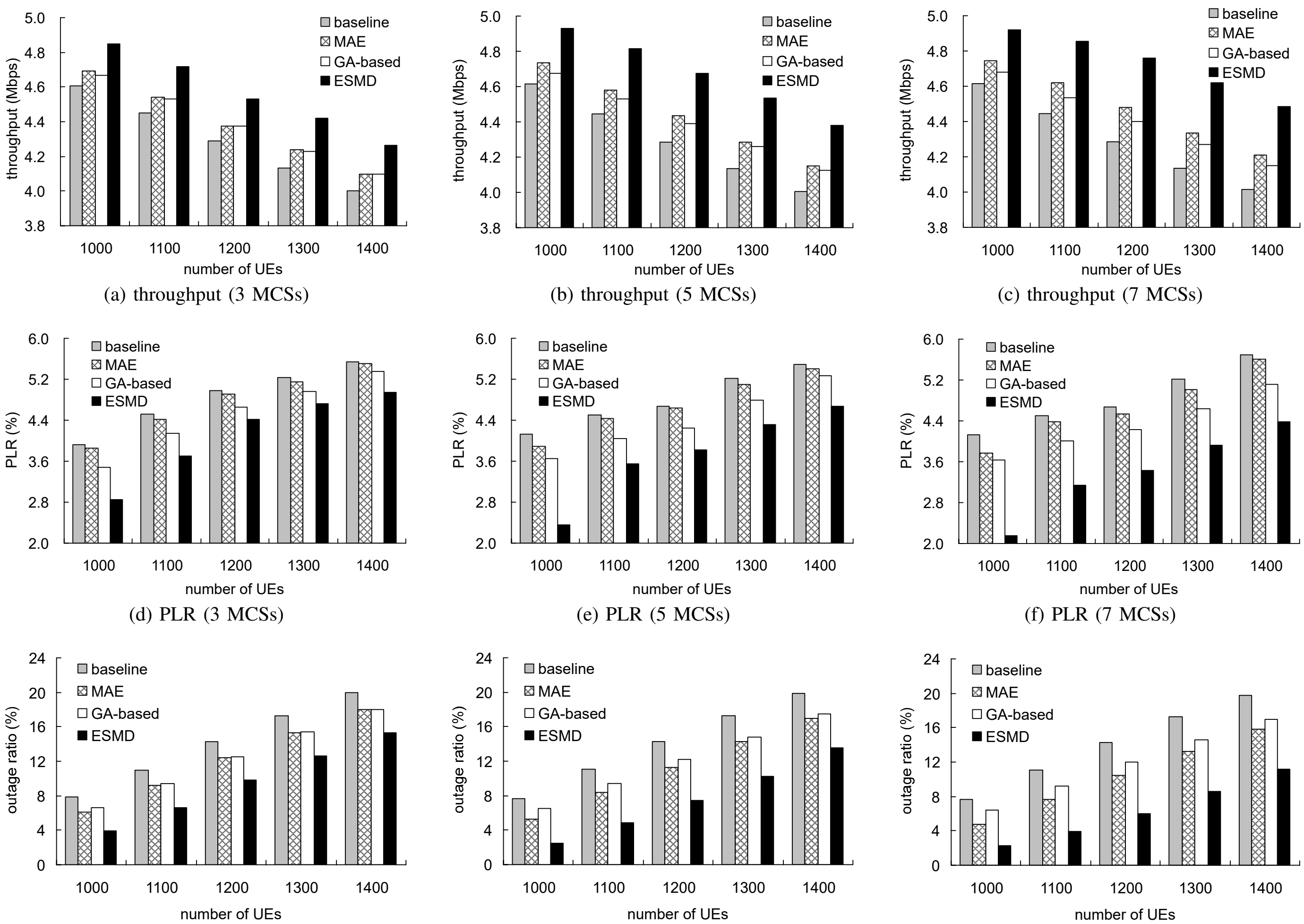
Proof: Each MCS uses a preference list to rank locations in $\hat{\mathcal{L}}$, which requires $O(\xi_L \log_2 \xi_L)$ time to do sorting. As there are ξ_M MCSs in $\hat{\mathcal{M}}$, it takes $O(\xi_M \times (\xi_L \log_2 \xi_L))$ time to build all preference lists. Based on the implementation in [22], repeated trade-in-free and coalition-free checks can be done by searching all preference lists twice, which spends $2O(\xi_M \times \xi_L)$ time. Thus, the time complexity is $O(\xi_M \times (\xi_L \log_2 \xi_L)) + 2O(\xi_M \times \xi_L) = O(\xi_M \times (\xi_L \log_2 \xi_L))$. ■

V. PERFORMANCE EVALUATION

We develop a simulator in MATLAB to evaluate the system performance. The service area is a 5 km × 5 km square, inside which 5G gNBs are deployed. A few MCSs move in the area to provide services. Table I gives the parameters of gNBs and MCSs. Moreover, we adopt the log-distance model to measure the amount of signal's attenuation caused by path loss [23]:

$$\text{gNB:} \quad 28 + 22\log_{10}(d[\text{m}]) + 20\log_{10}(f_c[\text{GHz}]), \tag{4}$$
$$\text{MCS:} \quad 32.4 + 21\log_{10}(d[\text{m}]) + 20\log_{10}(f_c[\text{GHz}]), \tag{5}$$

where d[m] is the distance from a gNB or MCS to a UE, which is measured in meters, and f_c[GHz] is the operating frequency band. Both gNBs and MCSs operate in the 28 GHz band. A zero-mean log-normal distribution is used to estimate the effect of shadowing fading. Its standard deviation is set to 10 dB and 6 dB for gNBs and MCSs, respectively. For the environmental noise, its power spectral density is -174 dBm/Hz [24].



(a) throughput (3 MCSs)
(b) throughput (5 MCSs)
(c) throughput (7 MCSs)
(d) PLR (3 MCSs)
(e) PLR (5 MCSs)
(f) PLR (7 MCSs)
(g) outage ratio (3 MCSs)
(h) outage ratio (5 MCSs)
(i) outage ratio (7 MCSs)

Fig. 2. Experimental results.

There are 1000, 1100, 1200, 1300, and 1400 UEs, where a half of them are randomly distributed over the service area. To simulate the crowded situation, we arbitrarily select five 1 km $\times$ 1 km regions to be hotspots. Around 10% of UEs congregate in each hotspot. The traffic demand of each UE is 5 Mbps.

We compare our ESMD scheme with three methods. First, the *baseline* method uses only gNBs to serve UEs, which helps assess the effect of using MCSs on performance. Second, the *maximizing service (MAE)* method chooses a location for each MCS such that it can cover the maximum UEs. Besides, the distance between two neighboring MCSs is more than 100 m to prevent them from interfering with each other. Third, the GA-based method [14] dispatches MCSs by a genetic algorithm, as mentioned earlier in Section II. For every experiment, we repeat simulations 100 times and take their average.

Fig. 2(a)–(c) give the average throughput of UEs by using 3, 5, and 7 MCSs. As UEs vie for the fixed resources, more UEs lead to lower throughput. Since the baseline method uses only gNBs to serve UEs, changing the number of MCSs will not affect its performance. On the other hand, the throughput in other three methods can rise by adding more MCSs. This result shows the benefit of using MCSs. Moreover, the MAE method asks each MCS to cover the maximum number of UEs, so it has higher throughput than the GA-based method. Our ESMD scheme can efficiently dispatch MCSs to serve UEs, so it always keeps the highest throughput.

Fig. 2(d)–(f) present the *packet loss rates (PLRs)* of UEs with different numbers of MCSs. Evidently, the PLR increases when there are more UEs. Besides, the PLR in each method (except the baseline method) can reduce by using more MCSs. Interestingly, the GA-based method has smaller PLRs than the MAE method (despite its throughput is lower than MAE). It means that the GA-based method considers packet latency when dispatching MCSs. By assigning MCSs to move to the selected locations based on a Pareto-optimal matching, ESMD can greatly reduce the PLR, as compared with other methods.

By changing the number of MCSs, Fig. 2(g)–(i) show the outage ratio, which is defined by the percentage of UEs that are not given RBs. Without the MCS's help, the baseline method has the largest outage ratio. Since MAE dispatches MCSs to

TABLE II

IMPROVEMENT RATIOS BY OUR ESMD SCHEME AS COMPARED WITH OTHER METHODS.

	average throughput			**PLR**			**outage ratio**		
method	3 MCSs	5 MCSs	7 MCSs	3 MCSs	5 MCSs	7 MCSs	3 MCSs	5 MCSs	7 MCSs
baseline	6.0%	8.7%	10.0%	15.4%	22.8%	30.4%	33.9%	48.7%	57.5%
MAE	3.8%	5.2%	5.6%	14.1%	21.0%	27.7%	23.3%	35.3%	41.8%
GA-based	4.0%	6.2%	7.3%	9.3%	15.8%	22.0%	24.7%	40.3%	49.7%

cover the most UEs, it can lower the outage ratio, as compared with the GA-based method. Thanks to the eAHC strategy used in the scheduling module, our ESMD scheme can efficiently find out candidate locations to place MCSs to serve NSY UEs, thereby substantially decreasing the outage ratio.

Table II lists the improvement ratios by the ESMD scheme, as compared with the baseline, MAE, and GA-based methods. As can be seen, our proposed ESMD scheme outperforms all other methods, especially in the outage ratio. Moreover, ESMD can improve the system performance more efficiently as there are more MCSs in the service area.

VI. CONCLUSION

When many UEs assemble in a hotspot region, the gNBs in the region may not have enough RBs to serve them, which degrades performance. The problem can be solved by adopting MCSs to flexibly move to provide services to these UEs. In this paper, we propose the ESMD scheme to help gNBs allot RBs to UEs and find candidate locations to place MCSs through the eAHC strategy. By finding a Pareto-optimal matching, ESMD well dispatches MCSs to move to serve NSY UEs. Simulation results show that ESMD can efficiently increase throughput, decrease the PLR, and reduce the outage ratio, as compared with the baseline, MAE, and GA-based methods. For future work, we will consider that flows have different priorities [25]. In this case, it is important to schedule resources and dispatch MCSs to satisfy the QoS demands of high-priority flows while preventing low-priority flows from starvation.

ACKNOWLEDGMENT

This work was supported by the Ministry of Science and Technology, Taiwan under Grant 108-2221-E-110-016-MY3.

REFERENCES

[1] F. Rinaldi, A. Raschella, and S. Pizzi, "5G NR system design: A concise survey of key features and capabilities," *Wireless Networks*, vol. 27, pp. 5173–5188, 2021.

[2] Cisco, "Annual Internet Report (2018–2023)," https://www.cisco.com/.

[3] Y. C. Wang and C. A. Chuang, "Efficient eNB deployment strategy for heterogeneous cells in 4G LTE systems," *Computer Networks*, vol. 79, pp. 297–312, 2015.

[4] S. F. Chou, Y. J. Yu, and A. C. Pang, "Mobile small cell deployment for service time maximization over next-generation cellular networks," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 6, pp. 5398–5408, 2017.

[5] Z. Becvar, M. Vondra, P. Mach, J. Plachy, and D. Gesbert, "Performance of mobile networks with UAVs: Can flying base stations substitute ultra-dense small cells?" in *European Wireless Conference*, 2017, pp. 261–267.

[6] H. Shinbo, Y. Kunisawa, T. Sakai, Y. Kitatsuji, A. Endo, and K. Tanaka, "Flying base station for temporary mobile communications in an area affected by a disaster," in *International Conference on Information and Communication Technologies for Disaster Management*, 2018, pp. 1–7.

[7] J. Lyu, Y. Zeng, R. Zhang, and T. J. Lim, "Placement optimization of UAV-mounted mobile base stations," *IEEE Communications Letters*, vol. 21, no. 3, pp. 604– 607, 2017.

[8] A. Fotouhi, "Towards intelligent flying base stations in future wireless network," in *IEEE International Symposium on A World of Wireless, Mobile and Multimedia Networks*, 2017, pp. 1–3.

[9] M. Huang, L. Huang, S. Zhong, and P. Zhang, "UAV-mounted mobile base station placement via sparse recovery," *IEEE Access*, vol. 8, pp. 71 775–71 781, 2020.

[10] M. D. Ree, G. Mantas, J. Rodriguez, and I. E. Otung, "Distributed trusted authority-based key management for beyond 5G network coding-enabled mobile small cells," in *IEEE 5G World Forum*, 2019, pp. 80–85.

[11] J. Chen and D. Gesbert, "Optimal positioning of flying relays for wireless networks: A LOS map approach," in *IEEE International Conference on Communications*, 2017, pp. 1–6.

[12] M. Alzenad, A. El-Keyi, F. Lagum, and H. Yanikomeroglu, "3-D placement of an unmanned aerial vehicle base station (UAV-BS) for energy-efficient maximal coverage," *IEEE Wireless Communications Letters*, vol. 6, no. 4, pp. 434–437, 2017.

[13] M. M. Azari, F. Rosas, K. C. Chen, and S. Pollin, "Ultra reliable UAV communication using altitude and cooperation diversity," *IEEE Transactions on Communications*, vol. 66, no. 1, pp. 330–344, 2018.

[14] Z. Becvar, P. Mach, J. Plachy, and M. F. P. de Tudela, "Positioning of flying base stations to optimize throughput and energy consumption of mobile devices," in *IEEE Vehicular Technology Conference*, 2019, pp. 1–7.

[15] Y. C. Wang and D. R. Jhong, "Efficient allocation of LTE downlink spectral resource to improve fairness and throughput," *International Journal of Communication Systems*, vol. 30, no. 14, pp. 1–13, 2017.

[16] Y. C. Wang and S. T. Chen, "Delay-aware ABS adjustment to support QoS for real-time traffic in LTE-A HetNet," *IEEE Wireless Communications Letters*, vol. 6, no. 5, pp. 590–593, 2017.

[17] A. K. Thyagarajan, P. Balasubramanian, D. Vydeki, and M. Karthik, "SNR-CQI mapping for 5G downlink network," in *IEEE Asia Pacific Conference on Wireless and Mobile*, 2021, pp. 173–177.

[18] Y. C. Wang and T. Y. Tsai, "A pricing-aware resource scheduling framework for LTE networks," *IEEE/ACM Transactions on Networking*, vol. 25, no. 3, pp. 1445–1458, 2017.

[19] Y. C. Wang and S. J. Liu, "Minimum-cost deployment of adjustable readers to provide complete coverage of tags in RFID systems," *Journal of Systems and Software*, vol. 134, pp. 228–241, 2017.

[20] J. E. Hopcroft and R. M. Karp, "A $n^{5/2}$ algorithm for maximum matchings in bipartite graphs," *SIAM Journal on Computing*, vol. 2, pp. 225–231, 1973.

[21] Y. C. Wang, "A two-phase dispatch heuristic to schedule the movement of multi-attribute mobile sensors in a hybrid wireless sensor network," *IEEE Transactions on Mobile Computing*, vol. 13, no. 4, pp. 709–722, 2014.

[22] D. J. Abraham, K. Cechlarova, D. F. Manlove, and K. Mehlhorn, "Pareto optimality in house allocation problems," *Proc. International Conference on Algorithms and Computation*, pp. 1163–1175, 2005.

[23] ETSI, "Study on channel model for frequencies from 0.5 to 100 GHz," 3GPP TR 38.901 V16.1.0 (Release 16), Nov. 2020.

[24] Y. C. Wang and S. Lee, "Small-cell planning in LTE HetNet to improve energy efficiency," *International Journal of Communication Systems*, vol. 31, no. 5, pp. 1–18, 2018.

[25] Y. C. Wang and K. C. Chien, "EPS: energy-efficient pricing and resource scheduling in LTE-A heterogeneous networks," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 9, pp. 8832–8845, 2018.