Service-Based Resource Scheduling Optimization for Multi-User OTFS-Based Systems

Ahmad M. Jaradat, Mohanad Alayedi, and Hüseyin Arslan, *Fellow, IEEE*

***Abstract*—In this letter, we investigate a flexible resource allocation (RA) in the delay and Doppler domains to accommodate services with varying requirements. We demonstrate that our problem is NP-hard and propose an optimization method based on linear programming (LP) and Lagrangian dual (LD). The adoption of flexibility in the delay and Doppler domains for capacity enhancement and addressing the needs of mission-critical services has considerable advantages, according to the obtained numerical results. They show that the new scheduling method outperforms conventional time-frequency (TF) domains in terms of bit rate and latency per user, making it valuable for fifth generation (5G) applications and beyond.**

***Index Terms*—Frame structure, 5G NR, scheduling, resource allocation, OTFS, resource block.**

I. Introduction

ORTHOGONAL Frequency Division Multiplexing (OFDM) is a widely adopted multi-carrier technique, extensively used in various IEEE standards. However, it experiences inter-carrier interference (ICI) due to high mobility, which leads to a reduction in channel capacity [1], [2]. Recently, Orthogonal Time Frequency Space (OTFS) modulation has been proposed as an alternative to OFDM because it has demonstrated greater resilience to Doppler spread compared to traditional OFDM. OTFS modulation maps information in the delay-Doppler (DD) domain, unlike OFDM modulation, which operates in the time-frequency (TF) domain [3]. Unlike OFDM technique, OTFS extends each information symbol in the DD domain throughout the whole TF domain, allowing for full TF diversity [3]. Because each information symbol sees the same constant channel gain, the transmitter and receiver designs are considerably simplified [4].

Nowadays, the need for using the fifth generation (5G) mobile communications is considerable increasing and has become inevitable to accommodate a broad range of services [5]. The scalable transmission time intervals (TTIs) appear to be a promising approach [6], [7]. Flexibility in resource optimization across both dimensions, termed as 2-D resource allocation (RA), presents new challenges [8], [9]. These challenges arise from the need to consider parameters such as channel conditions, number of users, and traffic types when developing a scheduling algorithm. Each user's request in the 2-D case must be assigned a 2-D resource units (RUs) allocation structure.

Resource scheduling in wireless communications has evolved significantly over the years, driven by the need to support increasing data rates, diverse service requirements, and higher user densities. Early methods focused on basic time and frequency domain allocation, while more recent approaches incorporate advanced techniques like network slicing and full-duplex communication to optimize resource use [10], [11].

Previous research has explored RA in 5G networks using linear programming (LP) and Lagrangian dual (LD) methods. LP has been effective in providing optimal solutions, such as in [12], where it efficiently allocated resources in the TF domain considering Quality of Service (QoS) requirements, and in [13], where it optimized downlink RA in multi-user Orthogonal Frequency Division Multiple Access (OFDMA) systems, improving throughput and fairness. LD methods have been studied for their effectiveness in handling large-scale optimization by decomposing problems into manageable sub-problems. For instance, [14] used an LD-based approach for joint power and resource allocation in heterogeneous networks, achieving near-optimal solutions with reduced complexity. Similarly, [15] developed an LD-based algorithm for dynamic spectrum access in cognitive radio networks, efficiently allocating resources while meeting interference constraints.

Service-based scheduling is crucial in modern wireless communication systems, especially with the diverse range of services offered by 5G networks [16]. Each service has unique requirements in terms of latency, bandwidth, and reliability. Efficiently allocating resources to meet these specific requirements ensures optimal network performance and user satisfaction. Integrating service-specific considerations into resource scheduling is essential for maximizing the potential of OTFS-based systems.

Since OTFS was proposed in 2017, there has been a noticeable lack of commercial and research initiatives focused on expanding OTFS resources, despite its significant potential and practical applications in signaling. More specifically, the current literature has not yet examined the allocation of blocks to users with varying service requirements within OTFS-based systems. Accordingly, this letter aims to address this gap and provide pertinent information to both academic and industry

Manuscript received 20 June 2024; revised 30 July 2024; accepted 1 August 2024. Date of publication 6 August 2024; date of current version 9 October 2024. The associate editor coordinating the review of this article and approving it for publication was G. Alexandropoulos. *(Corresponding author: Ahmad M. Jaradat.)*

Ahmad M. Jaradat is with the Department of Electrical and Computer Engineering and Computer Science, University of Detroit Mercy, Detroit, MI 48221 USA (e-mail: jaradaaam@udmercy.edu).

Mohanad Alayedi is with the Department of Software Engineering, Faculty of Engineering, Haliç University, 34060 Istanbul, Türkiye, and also with the Department of Electrical and Electronics Engineering, Istanbul Medipol University, 34810 Istanbul, Türkiye (e-mail: mohanad.alayedi@gmail.com).

Hüseyin Arslan is with the Department of Electrical and Electronics Engineering, Istanbul Medipol University, 34810 Istanbul, Türkiye (e-mail: huseyinarslan@medipol.edu.tr).

Digital Object Identifier 10.1109/LWC.2024.3438798

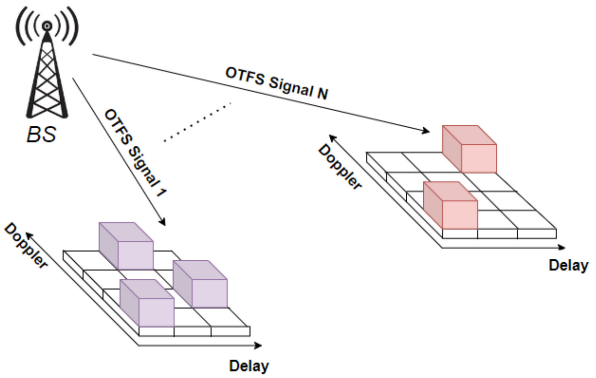

Fig. 1. The scheduling system model based OTFS signal.

sectors. To achieve this, we have developed a scheduling algorithm for OTFS-based wireless technology. The goal is to optimize the allocation of OTFS resources for block-service assignments. This algorithm leverages both LP and LD methods.

The contribution of this letter can be summarized as follows:

1) We tailor our optimization task to align with the requirements of 5G networks. In this context, we develop a new scheduling algorithm to support multiple services based on DD domains.
2) Demonstrate enhanced performance in terms of bit rate and latency per user for the OTFS waveform compared to the OFDM waveform.
3) The proposed algorithm enables the base station (BS) to dynamically communicate with multiple users to support multiple OTFS services.

The benefits of the proposed scheduling algorithm are outlined as follows:

- The implementation of the proposed scheduler does not need any modifications to the per-user scheduling policies used by BSs.
- We achieve an efficient and reliable performance when considering the DD domain.
- The versatility of the proposed scheduler enables its adaptation to various scenarios in Beyond 5G networks.

The rest of this letter is structured as follows: The second part details our proposed system model followed by problem formulation in the third part. The fourth part presents the proposed algorithm, while the fifth part discusses our simulation results. Finally, the sixth part concludes by summarizing the motivation behind this letter and suggesting future directions based on our findings.

II. System Model

We consider an OTFS system with a total length of T_f and a bandwidth of B. The bandwidth is divided into M subcarriers with δ_f subcarrier bandwidth. N symbols are transmitted, each with a symbol duration of T. Our system involves a BS providing two different services as illustrated in Figure 1. The first category, denoted by $S(n)$, has a tight latency constraint. Denote the data demand for any service $s \in S(n)$ as d_s (in bits), which must be met with a latency tolerance of l_s.

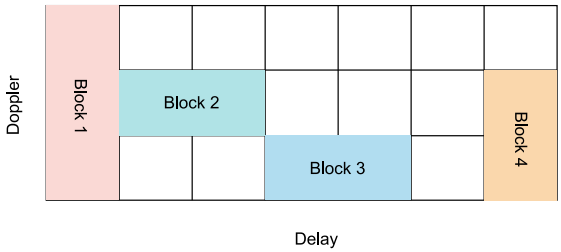

Fig. 2. An illustration of RA with four types of blocks. A rectangle of the grid is a basic unit of the resource. A service is allocated with one or multiple blocks, and each block can be assigned to up to one service.

Latency tolerance refers to the time it takes for the scheduler to fully transmit the data. To meet the entire delivery date, the parameter can be modified to account for queue delays as well as time spent at the receiver for processing/computing. The second type of service is denoted by $S(h)$, and the goal is to maximize throughput. Full buffer is assumed for services in $S(h)$. Furthermore, services in $S(n)$ are given precedence over those in $S(h)$. S is defined as $S(n) \cup S(h)$.

The number of slots within a frame defines the radio frame structure. A single RA to a service consists of a group of contiguous delay and Doppler RUs in the delay and Doppler domains, respectively. We refer to this group as a block, and consider a candidate set K of blocks, as shown in Figure 2. The achieved throughput on block k if k is assigned to service s ($s \in S$) is represented by $p_{k,s}$ for each $k \in K$. $p_{k,s}$ is dependent on the configuration of block k, including the time and frequency range (defined by subcarrier spacing (SCS) and TTI duration), and the symbol duration, given the channel profile, the transmission power, and the noise power. To compute the achieved throughput per block, we assume a total number of nine multipath channel profiles [17].

III. Problem Formulation

We consider the problem of maximizing total throughput for $S(h)$ while considering latency and demand limits for $S(n)$. In the problem formulation, the term "basic unit" refers to the smallest unit of resource in the DD domain. U stands for the set of basic units. Let's set $g_{k,u} = 1$ if basic unit u is present in block k; else, $g_{k,u} = 0$. Remember that a block is a resource shape in the resource grid. Arbitrary resource shapes increase the computational complexity of the optimization problem. Therefore, we consider rectangular resources to simplify the problem and make it more tractable, especially in real-time scenarios where computational resources are limited; for an example, see Figure 2. The rectangular block made up of RUs. In Figure 2, By putting the block 1 of shape 1×4 in each of the grid's conceivable locations, we can determine all of the blocks and their matching g-values.

This produces the collection of all candidate blocks K for all shapes. The position of the block and the numerical indexing of the basic units completely determine the g-values for each block, as shown in Figure 2, and the difficulty of performing one mapping equals the size of U, i.e., $|U|$. The complexity is $\mathcal{O}(|K||U|)$ to obtain all of the mappings. As it only concerns

the two sets K and U, this mapping is completed in the pre-processing stage once (not for each TTI). It is completely independent of service scheduling.

The goal of optimization is to choose blocks for each service that satisfy the latency and demand criteria for $S(n)$ without overlapping. For the purpose of determining if block k ($k \in K$) is assigned to service s ($s \in S$), we employ the optimization variable $m_{k,s} \in \{1, 0\}$. If the termination time of block k is longer than l_s, then k is not feasible for s ($s \in S(n)$). To simulate this, set $p_{k,s} = 0$. The problem is stated below.

$$\begin{aligned}[\text{P0}] \max_{\boldsymbol{m}\in\{0,1\}} \ & \sum_{k\in\mathcal{K}} \sum_{s\in\mathcal{S}^{(h)}} p_{k,s} m_{k,s} \\ \text{s.t.} \ & \sum_{k\in\mathcal{K}} p_{k,s} m_{k,s} \geq d_s, \quad s \in \mathcal{S}^{(n)} \\ & \sum_{s\in\mathcal{S}} \sum_{k\in\mathcal{K}} g_{k,u} m_{k,s} \leq 1, u \in \mathcal{U} \end{aligned} \tag{1}$$

The optimization problem P0 focuses on achieving block-service assignment that maximizes the overall throughput for $S(h)$ with two sets of constraints. The first constraint guarantees that the minimum data demand for $S(n)$, denoted as d_s, is satisfied within a latency tolerance of l_s. The second constraint ensures that the blocks do not overlap.

Theorem 1: P0 is NP-hard. Proof: For a group of integers $\{a_1, \ldots, a_b\}$, we create a polynomial-time reduction from the partition problem (PP) [18]. Assuming that $\acute{\alpha} = \sum_{k=1}^{b} a_k$ is even, the aim is to discover whether or not there is a partition such that the two subsets have the same sum of $\acute{\alpha}/2$. We define numerous blocks that form a single basic unit and a single TTI size. In $S(n)$ and $S(h)$, there are two services, indicated by s^n and s^h, respectively. The TTI size is the same as the latency parameter of s^n, and the demand is equal to $\acute{\alpha}/2$. Moreover, $p_{k,1} = p_{k,2} = d_c, c = 1, \ldots, b$. By construction, the second constraint in P0 and P0-LP has no impact. After that, it can be seen that partitioning the b basic units into two subsets, each of which offers an equal throughput of $\acute{\alpha}/2$, is comparable to finding a workable solution to PP. Furthermore, this is only possible if the objective function for s^h exceeds $\acute{\alpha}/2$. Hence the conclusion.

IV. Problem Solution

In this letter, we offer a sub-optimal yet low-complexity method that assigns blocks to services based on utility values produced through relaxation of LP and LD.

A. Assignment of Blocks

The matrix r with the dimensions $|K| \times |S|$ is the considered utility matrix for all pairs of blocks and services. The utility of a block-service pair (k, s) (with $k \in K$ and $s \in S$) is represented by an element $r_{k,s}$. Block assignments are handled differently for $S(n)$ and $S(h)$. Due to the requirement for latency, the former is carried out first. Let's assume F denotes the block-service assignment, indicating which blocks are already assigned to services. In the proposed Algorithm 1, the blocks that overlap with the blocks already assigned in F is removed. After ensuring non-overlapping blocks, the optimization problem P0 is determined as in (1). The proposed algorithm aims to estimate the values of $r_{k,s}$ using either LP or LD techniques. LP and LD can be used in conjunction, with LD helping to solve the problem more efficiently. After solving P0, check if the demands of the users left in $S(n)$ can be met. If the demands cannot be met, it suggests that the current assignment may not be feasible.

Algorithm 1 Proposed Scheduling Algorithm for Optimized Resource Allocation in OTFS-Based Systems

Require: Block-service assignment F and a utility matrix $r_{k,s}$. ▷ Note: k overlaps with k' if and only if $\exists u \in \mathcal{U}$ such that $g_{k,u} + g_{k',u} > 1$

1: **repeat**
2: Remove from K the blocks in F and those overlapping with the blocks in F.
3: $(k', s') \leftarrow \arg\max_{k\in K, s\in S(n)} u_{k,s}$, $F \leftarrow F \cup \{(k', s')\}$
4: **if** d_s is met **then**
5: $S(n) \leftarrow S(n) \setminus \{s'\}$
6: **end if**
7: **until** $S(n) = \emptyset$ or $K = \emptyset$
8: **if** $S(n) \neq \emptyset$ **then**
9: The demand of the users left in $S(n)$ cannot be met
10: **end if**
11: **repeat**
12: Repeat lines 2–3 with the notation $S(n)$ replaced by $S(h)$.
13: **until** $K = \emptyset$

We used the "Big O notation, i.e., $\mathcal{O}$" which is often used to measure the complexity of such algorithms, to calculate the complexity of each step in our algorithm. According to hash-map implementations, the complexity order of the operations in Lines 2 and 4 is $\mathcal{O}(1)$. The overall complexity order can be determined by sorting the utilities beforehand, which has a cost of $\mathcal{O}(|K||S|log(|K||S|))$ for both $S(n)$ and $S(h)$, and by looking at the other algorithm operations.

B. Utility Estimation by LP Relaxation

Solving the LP relaxation [19] of P0 using the LP optimal m_{LP} is one method of calculating r.

$$\begin{aligned}[\text{P0-LP}] \max_{0\leq \boldsymbol{m} \leq 1} \ & \sum_{k\in\mathcal{K}} \sum_{s\in\mathcal{S}(h)} p_{k,s} m_{k,s} \\ \text{s.t.} \ & \sum_{k\in\mathcal{K}} p_{k,s} m_{k,s} \geq d_s, \quad s \in \mathcal{S}^{(h)} \\ & \sum_{s\in\mathcal{S}} \sum_{k\in\mathcal{K}} g_{k,u} m_{k,s} \leq 1, u \in \mathcal{U} \end{aligned} \tag{2}$$

We define the LP-based utility as $r_{LP} = m_{LP}$. Also, m_{LP} can be used for initialization: $F = (k, s) : r_{LP,k,s} >= \beta, k \in K, s \in S$ with a threshold β.

C. Utility Estimation by LD

The Lagrangian is defined after relaxing the second constraint of (1) and (2) with the Lagrangian multiplier λ_u ($u \in \mathcal{U}$).

TABLE I
THE CONSIDERED SIMULATION PARAMETERS

Parameter	Value
Carrier frequency (f_c)	4 GHz
Number of candidate blocks (N_k)	2
Delay domain span (T)	66.667 μs
Doppler domain span (Δf)	15 kHz
Size of delay-Doppler resource block (DDRB) along delay domain (RU_T)	8.3337 μs
Size of DDRB along Doppler domain (RU_f)	1875
Demand (d)	4, 8, 12, 16 (kbps)
Latency tolerance (l)	0.25, 0.5, 1, 1.5, 2(ms)
Threshold (β)	0.05, 0.1,..., 0.95

$$Z(\boldsymbol{m},\boldsymbol{\lambda}) = \sum_{k\in\mathcal{K}}\sum_{s\in\mathcal{S}(h)} p_{k,s}m_{k,s} + \sum_{u\in\mathcal{U}}\lambda_u\left(1 - \sum_{k\in\mathcal{K}}\sum_{s\in\mathcal{S}} g_{k,u}m_{k,s}\right) \quad (3)$$

The LD function is defined in (3).

$$[\mathrm{P1}]\; v(\boldsymbol{\lambda}) = \max_{\boldsymbol{m}\in\{0,1\}} Z(\boldsymbol{m},\boldsymbol{\lambda})$$
$$\text{s.t.} \quad \sum_{k\in\mathcal{K}} p_{k,s}m_{k,s} \geq d_s, \quad s \in \mathcal{S}^{(h)} \quad (4)$$

Thus, we have the following LD problem:

$$[\mathrm{P0-LD}] \min_{\boldsymbol{\lambda}\geq 0} v(\boldsymbol{\lambda}). \quad (5)$$

Using a sub-gradient approach, the dual problem P0-LD can be resolved [20], [21], [22]. To solve the dual problem (5), we need to evaluate the dual function $g(\boldsymbol{\lambda})$ and compute its subgradient for any given dual variable (w). In the sub-gradient method, we start with an initial point $w^{(1)}$. At each iteration step $j = 1, 2, 3, \ldots,$ we compute the dual function, i.e., the objective function of the dual problem and a sub-gradient, then update w by

$$w^{(j+1)} = \left[w^{(j)} - o_j y^{(j)}\right]_+, \quad (6)$$

where $[.]_+$ denotes projection on the non-negative orthant, o_j is a positive scalar step-size, and y is the sub-gradient of the dual function. We used the simple step-size rule $o_j = t/j$, where t is a positive constant. The LD solution in the j^{th} iteration of the sub-gradient method is denoted as m^j_{LD}.

V. NUMERICAL RESULTS

It is anticipated that the flexible OTFS structure will perform better than non-flexible one. Examining the amount of improvement is the goal of performance evaluation, which is important in particular since the control channel overhead for maintaining the flexible structure is considered. The results also show how effectively the proposed algorithm fits into the flexible framework.

The considered OTFS system has a total length of 10 ms and a bandwidth of 10 MHz. The bandwidth is divided into 512 subcarriers with 15 kHz subcarrier bandwidth, and we transmit 128 symbols with a duration of $66.667\mu s$. We assume a total of ten users, with five requesting throughput maximization and the remaining five requesting delay tolerance. Table I lists the parameter settings. We test our method with a variety of possible thresholds ($\beta \in \{0.05, 0.5, \ldots, 0.95\}$), choosing the one that best achieves the goal. The configured number of sub-gradient iterations is 200. The locations of the considered blocks in the delay and Doppler domains are $[0 : 8.334\mu s, 0 : 1875]$ for the first block and $[8.334\mu s : 16.6680\mu s, 1875 : 3750]$ for the second block.

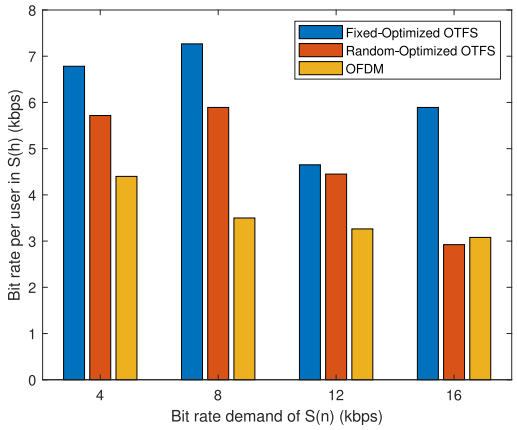

Fig. 3. Bit rate of $S(h)$ with respect to bit rate demand of $S(n)$ measured in kbps for OFDM and OTFS schemes.

In our study, we use both fixed and random OTFS-based schemes as benchmarks for comparison. These benchmarks are optimized versions of the fixed and random allocation methods. We compare these optimized benchmarks against the OFDM system using our optimization algorithm. The non-optimized fixed and random OTFS-based schemes represent common approaches in OTFS systems and provide suitable baselines for evaluating the effectiveness of our proposed method. The key difference between the non-optimized fixed and random OTFS schemes lies in their resource allocation strategies: fixed schemes allocate resources in a predetermined manner, while random schemes allocate resources through random assignment, without adhering to a specific pattern or order.

Figure 3 shows the bit rate of $S(h)$ with respect to bit rate demand of $S(n)$ measured in kbps. It shows that as the bit rate demand of user requesting $S(n)$ service increases, the bit rate of $S(h)$ also decreases due to requesting more data by the users for latency-constrained service. For simplicity, we have chosen two blocks and two services with one RU used in each block. In Figure 3, the optimal values of the optimization variable (m) to achieve these results are 0.0676 and 0.5676 between the first service and second block, and 0.4324 between second service and second block. Additionally, the bit rate per user in the fixed OTFS scheme exceeds that of the random OTFS scheme.

Figure 4 displays the latency of $S(h)$ relative to the bit rate demand of $S(n)$, measured in seconds. By visually representing the trade-offs between these parameters, it reveals that with an increase in the bit rate demand for the $S(n)$ service, the latency of $S(h)$ exhibits either a decrease or increase in OTFS and OFDM. Importantly, it is noteworthy that the latency observed in OFDM is higher than that in OTFS, as anticipated. Furthermore, the latency of the fixed OTFS scheme is less than the random OTFS scheme.

The flexible OTFS construction performs noticeably better than the non-flexible one. When the latency tolerance is

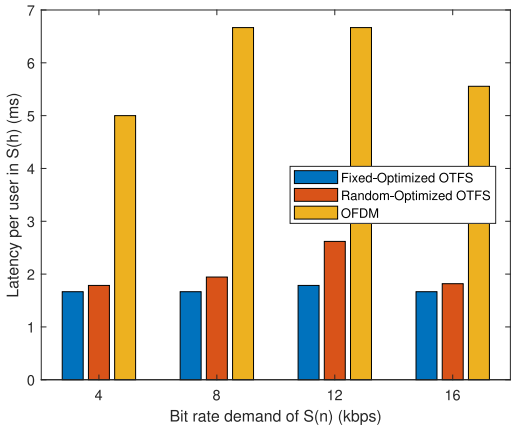

Fig. 4. Latency of $S(h)$ with respect to bit rate demand of $S(n)$ measured in seconds for OFDM and OTFS schemes.

tightened, the system often gains more from flexible structure. This result relates to the fact that, in the optimization problem, the throughput of the service $S(h)$ is subject to latency limitations of service $S(n)$. The former has few options for satisfying the latency specifications of $S(n)$, which results in low throughput for $S(h)$. Due to frequency selectivity, the latter is less effective in the frequency domain; however, when latency tolerance is large, this inefficiency is lessened because there are more options accessible in the time domain.

VI. Conclusion

We propose a flexible OTFS frame structure as a potentially effective choice for spectrum efficiency. Basically, our problem is determining how to assign 2-D blocks covering delay and Doppler to multiple users that request different services. This letter presents a novel scheduling algorithm for OTFS-based wireless technology that optimizes OTFS resources for block-service assignment. Our findings emphasize the need of optimizing block-service assignment. Our proposed algorithm was evaluated based on bit rate and latency per user and compared to OFDM. Results show that our scheduling algorithm outperforms OFDM in DD domains of OTFS modulation, unlike OFDM which relies on TF domains. As a potential future step, this letter could explore our scheduling approach's effectiveness in Orthogonal Delay-Doppler Division Multiplexing (ODDM), which shares similarities with the OTFS waveform in its reliance on DD domains. Moreover, it is crucial to examine the suitability of the scheduling algorithm in different areas, such as time-Fresnel domains, which are similar to Orthogonal Chirp Division Multiplexing (OCDM) waveform, in order to accomplish similar goals.

References

[1] A. M. Jaradat, J. M. Hamamreh, and H. Arslan, "OFDM with subcarrier number modulation," *IEEE Wireless Commun. Lett.*, vol. 7, no. 6, pp. 914–917, Dec. 2018.

[2] A. M. Jaradat, J. M. Hamamreh, and H. Arslan, "Modulation options for OFDM-based waveforms: Classification, comparison, and future directions," *IEEE Access*, vol. 7, pp. 17263–17278, 2019.

[3] R. Hadani et al., "Orthogonal time frequency space modulation," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, 2017, pp. 1–6.

[4] A. Farhang, A. RezazadehReyhani, L. E. Doyle, and B. Farhang-Boroujeny, "Low complexity modem structure for OFDM-based orthogonal time frequency space modulation," *IEEE Wireless Commun. Lett.*, vol. 7, no. 3, pp. 344–347, Jun. 2018.

[5] A. M. Jaradat, J. M. Hamamreh, and H. Arslan, "OFDM with hybrid number and index modulation," *IEEE Access*, vol. 8, pp. 55042–55053, 2020.

[6] A. M. Jaradat, A. Naeem, M. I. Sağlam, M. Kartal, and H. Arslan, "Radar-aided communication scheduling algorithm for 5G and beyond networks," *IEEE Access*, vol. 10, pp. 96403–96413, 2022.

[7] A. Anand, G. De Veciana, and S. Shakkottai, "Joint scheduling of URLLC and eMBB traffic in 5G wireless networks," *IEEE/ACM Trans. Netw.*, vol. 28, no. 2, pp. 477–490, Apr. 2020.

[8] A. M. Jaradat, M. I. Sağlam, M. Kartal, and H. Arslan, "Dynamic-structure resource block allocation based scheduling for 5G systems," in *Proc. IEEE 95th Veh. Technol. Conf. (VTC2022-Spring)*, 2022, pp. 1–5.

[9] Y. Ben-Shimol, I. Kitroser, and Y. Dinitz, "Two-dimensional mapping for wireless OFDMA systems," *IEEE Trans. Broadcast.*, vol. 52, no. 3, pp. 388–396, Sep. 2006.

[10] X. Foukas, G. Patounas, A. Elmokashfi, and M. K. Marina, "Network slicing in 5G: Survey and challenges," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 94–100, May 2017.

[11] Z. Zhang, K. Long, A. V. Vasilakos, and L. Hanzo, "Full-duplex wireless communications: Challenges, solutions, and future research directions," *Proc. IEEE*, vol. 104, no. 7, pp. 1369–1409, Jul. 2016.

[12] S. Ali et al., "Resource allocation in 5G networks: A survey," *IEEE Access*, vol. 7, pp. 9324–9363, 2019.

[13] A. Ahmad, A. Javaid, N. Javaid, M. Guizani, and A. Al-Fuqaha, "Fair and efficient resource allocation for 5G multi-user OFDMA systems using linear programming," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 11452–11466, Dec. 2018.

[14] X. Zhang, H. Dai, and R. Berry, "Joint power and resource allocation in heterogeneous networks using lagrangian dual methods," *IEEE Trans. Wireless Commun.*, vol. 15, no. 2, pp. 959–970, Feb. 2016.

[15] Q. Zhao, L. Tong, and A. Swami, "Dynamic spectrum access in cognitive radio networks using lagrangian dual decomposition," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 4, pp. 668–679, Apr. 2011.

[16] J. G. Andrews et al., "What will 5G be?" *IEEE J. Select. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[17] "Evolved universal terrestrial radio access (E-UTRA); user equipment (UE) radio transmission and reception, (Release 14)," 3GPP, Sophia Antipolis, France, Rep. TS 36.101, Apr. 2017. [Online]. Available: http://www.3gpp.org.

[18] J. A. Ruiz-Vanoye et al., "Survey of polynomial transformations between NP-complete problems," *J. Comput. Appl. Math.*, vol. 235, no. 16, pp. 4851–4865, 2011.

[19] K. Genova and V. Guliashki, "Linear integer programming methods and approaches–a survey," *J. Cybern. Inf. Technol.*, vol. 11, no. 1, PP. 1–23, 2011.

[20] S. Sen and H. D. Sherali, "A class of convergent primal-dual subgradient algorithms for decomposable convex programs," *Math. Program.*, vol. 35, no. 3, pp. 279–297, 1986.

[21] D. P. Bertsekas, "Nonlinear programming," *J. Oper. Res. Soc.*, vol. 48, no. 3, p. 334, 1997.

[22] N. Z. Shor, *Minimization Methods for Non-Differentiable Functions*, vol. 3. Berlin, Germany: Springer, 2012.