Completion-Time-Driven Scheduling for Uplink NOMA-Enabled Wireless Networks

Maryam Mohsenivatani, Ye Liu, Member, IEEE, Mahsa Derakhshani, Member, IEEE, Saeedeh Parsaeefard, Senior Member, IEEE, and Sangarapillai Lambotharan, Senior Member, IEEE

Abstract—Efficient scheduling policy is crucial in wireless networks due to delay-sensitivity of many emerging applications. In this work, we consider a joint user pairing and scheduling (UPaS) scheme for multi-carrier non-orthogonal multiple access (MC-NOMA)-enabled wireless networks to reduce the maximum completion time of serving uplink users. The NOMA scheduling problem is shown to be NP-hard and a shortest processing time (SPT)-based strategy to solve the same problem within affordable time and complexity is introduced. The simulation results confirm the efficacy of the proposed scheduling scheme in terms of the maximum completion time in comparison with orthogonal multiple access (OMA) and random NOMA pairing.

Index Terms—Non-Orthogonal Multiple Access (NOMA), maximum completion time, uplink scheduling.

I. INTRODUCTION

Non-orthogonal multiple access (NOMA) is one of the vital enabling technologies for the fifth generation (5G) of mobile networks and beyond. So far it has been mainly appreciated for its competence to dispense high data rates and substantial spectrum efficiency as a result of using successive interference cancellation (SIC) ([1], [2]). However, the potential of NOMA in latency reduction has not been fully explored.

When SIC is enabled, an uplink NOMA allows more than one transmitter to communicate to the same receiver simultaneously. This feature can be exploited to reduce the total required time to serve the uplink users. There have been some works (e.g., [3]–[6]) leveraging power control for a fixed user grouping to obtain minimal delay for single-channel NOMA-assisted mobile edge computing (MEC) networks. However, little effort has been spent on investigating the optimal scheduling and pairing to achieve transmission time reduction. The most broadly used criteria for delay optimality are the maximum completion time as well as the total completion time. The maximum completion time is the longest occupation time of all frequency channels and the total completion time is the sum of the completion times of all users. To the best of our knowledge, the work in [7] attempts to leverage NOMA and power control to reduce the maximum completion time. However, this work studies a system of only one frequency channel and the framework therein is not suitable for the multi-channel scenario that is ubiquitous nowadays.

This paper sheds light on the potentials of proper pairing and scheduling of uplink users in a multi-channel network to reduce the time required to serve every user. The work also considers the situation where different users transmit packets of different length, which may account for various delay constraints for different services [8]. In particular, we attempt to minimize the maximum completion time of the multi-user system, where the maximum completion time gives the longest busy time of all frequency channels. We formulate the optimization problem, establish the NP-hardness of the problem, and propose a NOMA-enabled scheduling in a multi-channel system that minimizes the maximum completion time. The proposed scheme saves 60% of time against OMA on average in the evaluated simulation settings.

II. SYSTEM MODEL AND PROBLEM FORMULATION

We consider uplink transmissions in which a base station (BS) at the center of a cell receives packets from $K$ users on $F$ frequency channels, each of size $W$ Hz. All terminals are single-antenna devices. Denote $K$ as the set of users. Each user $k$ has $L_k$ bits of data to send to the BS. We assume that at time zero, all users have their packets available for transmission. Additionally, we assume that each user $k \in K$ adapts its transmission power in such a way that the received signal-to-noise ratio (SNR) of the user at the BS on the $f$-th channel attains a specific value $\gamma_{k,f}$. A user is not allowed to transmit simultaneously through more than one frequency channel. Moreover, it is assumed that the received SNR of user $k$ has a specific value $\gamma_k$ regardless of which channel the user selects, i.e., $\gamma_{k,1} = \gamma_{k,2} = \cdots = \gamma_{k,F} = \gamma_k \ \forall k \in K$. Once a user started to transmit, it will be allowed to transmit the whole packet without interruption. Without loss of generality, we assume that $\gamma_1 \geq \gamma_2 \geq \cdots \geq \gamma_K$. Therefore, the achievable rate of user $k$ on any of the frequency channels in the OMA mode is

$$P_{k}^{\text{OMA}} = W \log_2(1 + h_k P_k / \sigma^2) = W \log_2(1 + \gamma_k).$$

where $P_k, h_k$, and $\sigma^2$ denote the transmit power of user $k$, the channel power gain between user $k$ and the BS, and the power...
of the noise at the receiver, respectively. Hence, the required transmission time of user \( k \) to finish its packet in the OMA mode is
\[
\tau_k = \frac{L_k}{R_k^\text{OMA}}. \tag{2}
\]

When SIC is enabled at the BS, we allow two users to transmit concurrently on any of the frequency channels. In this case, the BS will decode the user that has the highest SNR (say user \( i \)), remove its contribution to the received signal using SIC, and decode the weakest user signal (say user \( j \)). In this case, the achievable rate of user \( j \) is equal to \( R_j^\text{OMA} \), and the rate of user \( i \) is
\[
R_i^\text{SIC,j} = W \log_2\left(1 + \frac{h_i P_i}{h_j P_j + \sigma^2}\right) \tag{3}
\]
\[
= W \log_2\left(1 + \frac{h_i P_i}{h_j P_j + \sigma^2 + 1}\right) = W \log_2\left(1 + \frac{\gamma_i}{\gamma_j + 1}\right).
\]

Due to the interest in keeping the complexity of SIC low and also due to the advantages shown in the literature [9], we allow at most two users to access a frequency channel simultaneously in the NOMA mode as illustrated in Fig.1.

Then the required time for completing the transmission of user \( i \) when NOMA is applied for users \( i \) and \( j \) is
\[
\tau_i^\text{SIC,j} = \begin{cases} \tau_j, & \text{if } \frac{L_i}{R_i^\text{SIC,j}} \leq \frac{L_j}{R_j^\text{SIC,j}} \leq \frac{L_j}{R_j^\text{SIC,j}} \text{max}, \\ \tau_j + \frac{L_j - \tau_j}{\frac{R_j^\text{SIC,j}}{R_j^\text{SIC,j}}}, & \text{otherwise}. \end{cases} \tag{4}
\]

Let \( S \) denote a particular schedule and \( S \) denote the set of all possible schedules. The schedule \( S \) decides which task (i.e., which pair of users) will be scheduled at which channel for transmission. Let \( S = \{s_{i,j}\} \) where \( s_{i,j} \in \{0,1\} \) is the binary variable indicating whether the pair of users \( i \) and \( j \) has been scheduled for the channel \( f \) or not. The maximum completion time of a schedule \( S \) is defined as
\[
C_{\text{max}}(S) = \max_{f \in \mathcal{F}} \{C_f(S)\} \tag{5}
\]
where \( C_f(S) \) is the completion time of the last task assigned to channel \( f \) and is given by
\[
C_f(S) = \sum_{\forall i \in K} \sum_{\forall j \in K} s_{i,j} u_{i,j}, \forall f \in \mathcal{F}. \tag{6}
\]

For each task, in which users \( i \) and \( j \) (where \( i < j \in K \)) are paired to transmit with NOMA in channel \( f \) (i.e., \( s_{i,j}^f = 1 \)), \( u_{i,j} \) is the processing time which is represented as
\[
u_{i,j} = \max\{\tau_i^\text{SIC,j}, \tau_j\}. \tag{7}
\]

We consider the problem that the scheduler will minimize system’s maximum completion time \( C_{\text{max}}(S) \). We formulate the maximum completion time minimization problem with respect to schedule \( S \) as
\[
\min_{S \in \mathcal{S}} C_{\text{max}}(S), \text{ subject to } \sum_{\forall f \in \mathcal{F}} \sum_{\forall j \in K} s_{i,j}^f = 1, \forall i \in K, \quad s_{i,j}^f \in \{0,1\}. \tag{8}
\]

The first constraint ensures that each user can be scheduled only once. The above problem is NP-hard as established below.

## III. HARDNESS OF THE PROBLEM

In this section, we prove that the NOMA-enabled scheduling problem of minimizing \( C_{\text{max}} \) is NP-hard.

**Theorem 1.** When NOMA is enabled, the scheduling problem of minimizing \( C_{\text{max}} \) is NP-hard.

**Proof.** We prove this NP-hardness result by showing that a special case of the NOMA scheduling problem is equivalent to an NP-hard problem.

Assume that \( K \) is a multiple of two. Also, we assume the following relationship holds, i.e.,
\[
\frac{L_i}{R_i^\text{SIC,j}} \leq \frac{L_j}{R_j^\text{SIC,j}} \Rightarrow L_i \leq \frac{L_j}{R_j^\text{SIC,j}} L_j, \forall j > i. \tag{9}
\]

The relationship in (9) means that when the users \( i \) and \( j \) transmit simultaneously using NOMA, user \( i \) will always finish before user \( j \), where \( i < j \) such that the SIC order is first decode user \( i \) and then user \( j \). Notice that (9) implies that
\[
\tau_i < \tau_j, \forall i < j, \tag{10}
\]

because \( \tau_i = \frac{L_i}{R_i^\text{SIC,j}} \) and \( \frac{L_i}{R_i^\text{SIC,j}} \leq \frac{L_j}{R_j^\text{SIC,j}} \).

Given the assumption in (9) and together with the assumption that \( K \) is a multiple of two, we can perform \( K/2 \) NOMA transmissions to minimize \( C_{\text{max}} \). An optimal way to pair users using NOMA is \( N_2^1, N_4^1, N_2^2, \ldots, N_{K-1}^K \), where \( N_i^j \) denotes the NOMA transmission of users \( i \) and \( j \) such that \( i < j \). To see this, suppose \( N_a^b \) and \( N_d^c \) are part of the claimed optimal pairing, where \( a < b < c < d \), \( a = b - 1 \), and \( c = d - 1 \). Then, we can show that swapping the pairing can increase \( C_{\text{max}} \):

1. If we have \( N_a^b \) and \( N_d^c \) instead, then \( \tau_c \) and \( \tau_d \) would contribute to the time when all users are served, which can increase \( C_{\text{max}} \) compared to \( N_a^b \) and \( N_d^c \) because in the latter case, \( \tau_b \) and \( \tau_d \) contribute to \( C_{\text{max}} \), where \( \tau_b < \tau_c \) due to (10).
2. If we have \( N_a^d \) and \( N_b^c \), then \( C_{\text{max}} \) can be increased due to a similarity to the above case.
3. Any other swapping is not allowed because of the violation of SIC order.
Given the above optimal pairing, the NOMA scheduling problem reduces to an OMA scheduling problem, where we have equivalently \( K/2 \) users to serve and the time required to complete the OMA transmissions are \( \tau_2, \tau_3, \ldots, \tau_K \), respectively. The problem of minimizing \( C_{\max} \) for the OMA scheduling problem is exactly the same as the multiprocessor scheduling problem where \( K/2 \) jobs are to be processed by \( F \) identical machines, where the latter problem known to be NP-hard [10]. The above argument suggests that the NOMA scheduling problem belongs to the class of NP because it is equivalent to an NP-hard problem. It is apparent that the scheduling, when restricted by conditions in (9) and (10), is NP-hard [10]. The above argument suggests that the NOMA scheduling problem reduces to an OMA scheduling problem, where we can easily check whether a given schedule can meet the deadline \( C_{\max} \) or not. Therefore, (8) is NP-hard.

IV. PROPOSED UPaS ALGORITHM

Heuristics have been the centre of attention in the absence of optimal algorithms. Among the available scheduling heuristics for \( F \) identical machines and independent tasks (i.e., OMA), the shortest processing time (SPT) algorithm proves to be a viable option to obtain a minimal length schedule. This algorithm sorts the tasks in ascending order and starts scheduling the tasks on each machine from the shortest processing times and once a machine becomes available the task with the shortest processing time will be scheduled to transmit. Driven by the fact that for \( F \geq 1 \) identical processors SPT scheduling performs well in minimizing maximum completion time [11], a SPT-based joint UPaS for NOMA-enabled wireless networks is proposed.

We consider that all users have their packets ready for transmission at time zero. Assuming that the scheduler has access to perfect channel-state information (CSI) of all the links between the users and the BS, it first sorts the users in ascending order based on their required OMA transmission times. Given that we have \( K \) users and \( F \) channels, the scheduler clusters the users in \( M = \frac{K}{2} \) of size \( 2F \). Then, sequentially for each cluster, the scheduler schedules the first \( F \) users with the shortest processing times (i.e., OMA time) on the \( F \) available channels. Let \( U_m \) denote the set of \( F \) scheduled users belonging to cluster \( m \) and \( U'_m \) account for the remaining users in that cluster. Having considered NOMA, to find the best pair for the scheduled user on \( f \)th channel in cluster \( m \), the NOMA transmission times of the remaining users in \( U'_m \) is then calculated according to (4) considering that they should transmit along with the user already scheduled on the \( f \)th channel. Subsequently, the user with the shortest NOMA transmission time on that channel is then selected to transmit along with the already scheduled user.

Algorithm 1 summarizes the proposed scheme. Note that regarding the SIC order, the user with the higher received SNR value is decoded first.

<table>
<thead>
<tr>
<th>Algorithm 1: SPT-based joint user pairing and scheduling</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Result:</strong> (Pairs’ indices)</td>
</tr>
<tr>
<td>1 Cluster the users in ( K ) which are in ascending order ( (\tau_1 \leq \tau_2 \leq \cdots \leq \tau_K) ) into ( M = \frac{K}{2F} ) groups.</td>
</tr>
<tr>
<td>2 for ( m = 1 ) to ( M ) do</td>
</tr>
<tr>
<td>3 Assign the ( F ) first sorted users to the ( F ) available frequency channels.</td>
</tr>
<tr>
<td>4 for ( i \in U'_m ) do</td>
</tr>
<tr>
<td>5 for ( j \in U_m ) do</td>
</tr>
<tr>
<td>6 ( \tau_{ij} = \begin{cases} \frac{L_j}{R_{ij}^{\text{NOMA}}}, &amp; \text{if } \frac{L_j}{R_{ij}^{\text{NOMA}}} \leq \frac{L_j}{R_{ij}^{\text{OMA}}} \ \tau_j + \frac{L_j - R_{ij}^{\text{NOMA}}}{R_{ij}^{\text{OMA}}}, &amp; \text{otherwise} \end{cases} )</td>
</tr>
<tr>
<td>7 ( k_{m,i} = \arg\min_j \tau_{ij} )</td>
</tr>
<tr>
<td><strong>Output:</strong> ( (k_{m,i}, i) )</td>
</tr>
</tbody>
</table>

\( \text{(F+(F-1)+\cdots+1)} \) comparisons. Considering \( F = \frac{K}{2} \) the complexity of our proposed algorithm is as follows

\[ \frac{1}{4}(K^2 - K) + \frac{K}{2F}(\frac{K}{2} + (\frac{K}{2} - 1) + \cdots + 1) = O(K^2). \]

This shows the amount of comparisons needed with respect to the number of users in the cell.

B) Convergence Analysis - The approximation ratio of an algorithm to solve a combinatorial optimization problem is the ratio between the result obtained by the algorithm and the optimal value. Here we study the approximation ratio of our proposed algorithm. An algorithm with approximation ratio \( l \) is called an \( l \)-approximation algorithm; it is proved that any list scheduling (LS) algorithm[1] is 2-approximation for minimizing maximum completion time on \( m \) identical parallel machines [12]. This means

\[ C_{\max} \leq 2C_{\max}^* \]

Based on this fact, for a given pairs of users, it can be concluded that

\[ C_{\max}^{\text{NOMA}} \leq 2C_{\max}^{*,\text{NOMA}}, \quad (11) \]

where \( C_{\max}^{*,\text{NOMA}} \) is the optimal maximum completion time for the same given grouping. Since the completion time of group \( \{ \text{user}_j, \text{user}_i \} \) using NOMA is smaller than \( \tau_j + \tau_i \), therefore

\[ C_{\max}^{*,\text{NOMA}} < C_{\max}^{*,\text{OMA}}, \quad (12) \]

where \( C_{\max}^{*,\text{OMA}} \) is the optimal maximum completion time for OMA. Based on (11) and (12), the conclusion is that

\[ C_{\max}^{\text{NOMA}} < 2C_{\max}^{*,\text{OMA}} \]

1LS scheduling is a generic greedy algorithm, in which the tasks can be in any arbitrary order and any unscheduled task in this order can be assigned once a machine (channel here) becomes available.
were served by $F = 5$ channels. As shown, our proposed SPT-based joint UPaS outperforms other approaches in terms of both metrics. Comparing to OMA, our method has the ability to unlock the significant potential of NOMA as the service demand increases. The larger the number of users gets, the better our algorithm performs. By way of illustration, the performance improvement of our algorithm compared to OMA in terms of both the maximum completion time and the total completion time for $K = 20$ is almost 45% yet for $K = 50$ this improvement is almost 75%, which corroborates the potential of our proposed algorithms when the network experiences huge demand.

For a fixed number of users $K = 40$, as expected, Fig. 4 and Fig. 5, reveal that increasing the number of channels leads to a plunge in both the max and the total completion times. It appears that after reaching a certain threshold, NOMA does not seem to yield any significant benefit. NOMA is able to achieve the best it can within the limited availability of spectral resources. It is worth mentioning that it has been assumed that each user is supposed to transmit up to a certain data length. In Fig. 6, the effects of this data length limit on the maximum completion time have been illustrated. As observed, the proposed scheme is able to contribute to the performance by almost 75% as compared to the OMA counterpart. Fig. 7. gives insight about the fact that how the tasks are distributed to a plunge in both the max and the total completion times.

The maximum task completion time and the total completion time versus the number of users are depicted in Fig. 2 and Fig. 3 respectively for three different scenarios. The users

V. NUMERICAL RESULTS

In this section, the performance of our proposed joint SPT-based UPaS algorithm in terms of maximum completion time and total completion time of the uplink users is evaluated and compared to random NOMA pairing and scheduling, and OMA scheduling. In random NOMA pairing and scheduling, once the users are clustered and the first $F$ users are scheduled, the scheduler schedules a random pair for the already scheduled user on channel $f$. Note that RAND in the figures refers to the random NOMA pairing. The SPT algorithm is also used for OMA scheduling. In the simulations, users are randomly located within a circular area according to a uniform distribution with pathloss of $\beta_k = (d_k)^{-\zeta}$ where $d_k$ is the distance between user $k$ and the BS and $\zeta = 3$. It is assumed the packet size of every user is randomly chosen in the range of 1 bit to 1 Mbits. The bandwidth of each frequency channel is 1 MHz. Moreover, we have considered dynamic power control strategy in which far users transmit at relatively higher power and near users transmit at relatively lower power proportional to their average channel gains. To allocate power according to this strategy, we have considered that the users are placed such that the user $k$ is allocated $P_k = P_b \rho^{K-k}$. It means the farthest user is allocated $P_K = P_b \rho^{K-K} = P_b$, which we call the power budget. Similarly, the nearest user will be allocated $P_1 = P_b \rho^{K-1}$. The $\rho$ parameter needs to be selected according to the pathloss model such each user attains a satisfactory SNR. In our simulations, we chose $\rho$ as 0.9.

The maximum task completion time and the total completion time versus the number of users are depicted in Fig. 2 and Fig. 3 respectively for three different scenarios. The users

In Fig. 8, the performance difference between the proposed scheme and the baseline schemes in low-SNR and high-SNR regimes are presented. As shown the proposed algorithm is
considerably efficient in both low-SNR and high-SNR regimes and reduces the maximum completion time by 65% in comparison with OMA in the evaluated simulation settings.

In Fig.9, we have explored further and considered SIC error which led to the increase of the maximum completion time. In an uplink NOMA cluster, in a more realistic setting the BS cannot perfectly decode the weakest user signal (say user j). In this case, the achievable rate of user j according to model in [13] is equal to

\[ R_{j}^{\text{SIC},j} = W \log_2(1 + \frac{\gamma_j}{\beta \gamma_i + 1}), \]

where 0 < \beta < 1 is the coefficient of imperfect SIC, and the higher \beta means more interference because of the SIC error propagation [13]. The rate of user i with SIC (i.e., \( R_{i}^{\text{SIC},j} \)) will stay the same as in (3). The value of \beta is 0.1.

VI. CONCLUSION

We studied the problem of minimizing the maximum completion time of multiple uplink users that attempt to transmit packets of different lengths in a MC-NOMA system. We proved that this optimization problem is NP-hard, and we proposed a heuristic algorithm for NOMA scheduling with a complexity of \( O(K^2) \). The proposed algorithm for NOMA outperforms OMA in various scenarios not only in terms of the maximum completion time but also for the total completion time.

REFERENCES


