SUM-RATE MAXIMIZATION IN D2D-UNDERLAID NOMA CELL-FREE MEC NETWORKS VIA PARTIAL OFFLOADING AND SCA

TỐI ĐA HÓA TỔNG TỐC ĐỘ DỮ LIỆU TRONG MẠNG NOMA CELL-FREE MEC CÓ LỚP NỀN D2D THÔNG QUA OFFLOADING MỘT PHẦN VÀ PHƯƠNG PHÁP SCA

Tien V. Thai*, Mai T.P. Le, Hieu V. Nguyen, Hanh T.M. Tran

The University of Danang - University of Science and Technology, Vietnam

*Corresponding author: thaivantien@dut.udn.vn

(Received: April 24, 2026; Revised: July 02, 2026; Accepted: July 15, 2026)

DOI: 10.31130/ud-jst.2026.24(7A).255

Abstract - Integrating mobile edge computing (MEC) and non-orthogonal multiple access (NOMA) into cell-free massive MIMO (CF-mMIMO) networks is a potential approach for ultra-reliable and low-latency communication (URLLC) 6G. However, in dense deployments, underlay device-to-device (D2D) communication introduces heterogeneous spatial interference, rendering successive interference cancellation (SIC) based on channel gain and degrading NOMA performance. This paper examines resource allocation and partial computational offloading to maximize the total NOMA-CF-mMIMO MEC sum rate under latency and power constraints, and proposes a flexible offloading framework with SIC criteria based on the effective channel-to-interference ratio (ECIR). The NP-hard non-convex problem is solved using an alternating optimization (AO) algorithm combining ECIR-based NOMA-SIC pairing and successive convex approximation (SCA). Simulations show that the proposed method converges rapidly and consistently achieves a superior sum rate over conventional benchmarks, even under dense D2D interference.

Key words - Cell-free massive MIMO; mobile edge computing; non-orthogonal multiple access; D2D interference; partial offloading; successive convex approximation.

Tóm tắt - Tích hợp điện toán biên di động (MEC) và đa truy nhập phi trực giao (NOMA) vào mạng MIMO quy mô lớn phi tế bào (CF-mMIMO) là hướng tiềm năng cho truyền thông siêu tin cậy độ trễ thấp (URLLC) 6G. Tuy nhiên, trong triển khai dày đặc, truyền thông thiết bị–đến–thiết bị (D2D) dạng lớp nền gây nhiễu không gian không đồng nhất, làm vô hiệu hóa điều kiện triệt nhiễu nối tiếp (SIC) dựa trên độ lợi kênh và suy giảm hiệu năng NOMA. Bài báo xét cấp phát tài nguyên và giảm tải tính toán một phần nhằm tối đa hóa tổng tốc độ dưới ràng buộc độ trễ và công suất, với khung giảm tải linh hoạt và tiêu chí SIC dựa trên tỷ số kênh trên nhiễu hiệu dụng (ECIR). Bài toán phi lồi NP-khó được giải bằng thuật toán tối ưu hóa luân phiên (AO) kết hợp ghép cặp theo ECIR và xấp xỉ lồi liên tiếp (SCA). Mô phỏng cho thấy phương pháp đề xuất hội tụ nhanh và đạt tổng tốc độ vượt trội so với các phương pháp tham chiếu, kể cả khi nhiễu D2D dày đặc.

Từ khóa - MIMO quy mô lớn phi tế bào; điện toán biên di động; đa truy nhập phi trực giao; nhiễu D2D; giảm tải một phần; xấp xỉ lồi liên tiếp.

1. Introduction

The proliferation of computation-intensive and delay-sensitive applications - such as augmented reality (AR), autonomous vehicles, and the Industrial Internet of Things (IIoT) - is imposing unprecedented challenges on sixth-generation (6G) wireless networks. In addition to the requirements for massive connectivity and ultra-high spectral efficiency, these applications also demand extremely stringent end-to-end latency. In this context, mobile edge computing (MEC) has emerged as an important architectural advancement: by bringing cloud-like computing resources to the network edge, MEC enables computation offloading for user equipment (UEs), thereby significantly reducing transmission latency as well as UE energy consumption [1], [2], [3].

In parallel with MEC, the cell-free massive multiple-input multiple-output (CF-mMIMO) architecture is regarded as a breakthrough at the physical layer, enabling the inherent cell-boundary limitations of conventional cellular architectures to be overcome. In CF-mMIMO, a large number of geographically distributed access points (APs) cooperate with one another through high-capacity fronthaul links to jointly serve multiple UEs over the same time–frequency resource block.

By exploiting favorable propagation conditions and macro-diversity, CF-mMIMO can effectively suppress inter-cell interference and provide uniform quality of service (QoS) to all UEs [4]. This characteristic makes CF-mMIMO a particularly suitable access architecture for MEC deployment [5], [6].

1.1. Background and motivation

To expand massive connectivity capability and enhance spectral efficiency in the CF-mMIMO architecture, non-orthogonal multiple access (NOMA) has attracted considerable research attention [7], [8]. By combining superposition coding at the transmitter with successive interference cancellation (SIC) at the receiver, NOMA allows multiple UEs to share the same radio resource block. In addition, in practical 6G networks, device-to-device (D2D) communications are often deployed as an underlay spectrum-sharing layer to exploit spectrum resources more efficiently and enhance connectivity capability [9], [10].

The underlay deployment of D2D communications creates local and spatially heterogeneous interference at the APs of the CF-mMIMO network. In single-carrier NOMA-CF-mMIMO systems - where all NOMA clusters share the same frequency band - the conventional SIC ordering strategy, which relies solely on large-scale fading or unconditional channel gain [11], is no longer suitable in a spatially varying D2D interference environment. A UE with a very favorable channel to the cooperating APs may still experience strong local D2D interference, leading to SIC decoding errors and severe error propagation [12]. Furthermore, when the uplink becomes congested due to D2D interference, forcing UEs to offload their entire computational tasks to the MEC server inevitably leads to violations of latency constraints [13], [14]. Therefore, simultaneously managing D2D interference and satisfying stringent MEC latency constraints in NOMA-CF-mMIMO systems remains an open and critical problem.

1.2. Related work and research gap

The integration of MEC and CF-mMIMO has recently been actively investigated. Mukherjee and Lee [5] proposed an edge-computing-enabled CF-mMIMO framework and analyzed the probability of successful edge computation using stochastic geometry tools. Interdonato and Buzzi [6] examined the joint allocation of uplink transmit power and remote computing resources through multi-objective optimization to balance energy consumption and spectral efficiency. The latency–energy trade-off in MEC-enabled NOMA-CF-mMIMO systems was investigated in more detail in our previous work [14], and a Greedy-NOMA pairing strategy for multicarrier NOMA-CF-mMIMO was proposed in [15]. Along the direction of NOMA integration, Bashar et al. [7] studied an adaptive NOMA/OMA mode-switching mechanism in CF-mMIMO, while Dang et al. [16] proposed a user-pairing strategy based on inner approximation for NOMA-CF-mMIMO systems. However, most of these studies assume an interference-free environment from secondary networks and mainly target latency minimization or power minimization rather than sum-rate maximization.

In another research direction, resource allocation in D2D spectrum-sharing NOMA cellular networks was investigated in [17], and pilot reuse strategies for D2D-underlaid massive MIMO were proposed in [12]. However, these studies are limited to co-located antenna structures and therefore do not reflect the distributed signal reception nature of CF-mMIMO, nor do they consider the stringent latency constraints imposed by MEC. To the best of the authors' knowledge, the problem of combining partial computation offloading with D2D-interfered NOMA-CF-mMIMO - especially the design of an SIC criterion that explicitly accounts for spatial D2D interference - remains a significant research gap.

1.3. Contributions of this paper

This paper proposes a joint optimization framework for radio resource allocation and partial computation offloading in a single-carrier NOMA-CF-mMIMO network in the presence of underlay D2D links. The main contributions are summarized as follows:

- **Novel problem formulation:** We formulate a sum-rate maximization problem that can adapt to D2D interference by adjusting the offloading ratio: when D2D interference degrades the uplink capacity, the system proactively increases the local computation ratio to avoid uplink congestion and satisfy stringent latency constraints.
- **ECIR-based SIC ordering:** We introduce the effective channel-to-interference ratio (ECIR), which is specifically designed for single-carrier CF-mMIMO. This parameter jointly incorporates the cooperative beamforming gain after maximum-ratio combining (MRC) and the spatial distribution characteristics of heterogeneous D2D interference, thereby providing a robust SIC ordering criterion that faithfully reflects the actual reception conditions at the CPU. This is a fundamental distinction from conventional methods that rely solely on channel gain.
- **Efficient AO-SCA algorithm:** To solve the highly complex and non-convex mixed-integer nonlinear programming (MINLP) problem, we propose an alternating optimization (AO) algorithm combined with successive convex approximation (SCA) [9], which enables the original problem to be decomposed into tractable subproblems and guarantees monotonic convergence.
- **Complexity analysis and numerical validation:** We provide a step-by-step computational complexity analysis of the algorithm. Monte Carlo simulation results show that the proposed ECIR-NOMA technique achieves a stable sum-rate improvement of approximately 1.0–1.8% compared with the Random-NOMA and No-NOMA benchmark systems, while converging to 99% of the steady-state sum-rate value within only 10–15 iterations.

1.4. Paper organization and notation

The remainder of this paper is organized as follows. Section 2 presents the system model. Section 3 formulates the optimization problem. Section 4 details the proposed AO-SCA algorithm. Section 5 discusses the simulation results, and Section 6 concludes the paper.

Notation: Bold lowercase and uppercase letters denote vectors and matrices, respectively; $(\cdot)^T$ and $(\cdot)^H$ denote the transpose and conjugate transpose, respectively; $\|\cdot\|$ represents the Euclidean norm; $\mathcal{CN}(0,\sigma^2)$ denotes a circularly symmetric complex Gaussian distribution with zero mean and variance σ^2; $|\cdot|$ denotes the absolute value of a complex scalar; $\mathbb{E}[\cdot]$ is the expectation operator; $\mathbf{1}\{\cdot\}$ is the indicator function.

2. System model

2.1. Network architecture and channel model

We consider the uplink of a single-carrier CF-mMIMO MEC system consisting of M distributed single-antenna APs and K single-antenna UEs, with $M \gg K$ as illustrated in Figure 1. We assume that the APs are connected to the CPU through ultra-wideband optical fiber links; hence, the fronthaul latency is negligible compared with the wireless transmission latency and computational processing latency. An underlay layer consisting of D D2D transmitter–receiver pairs shares the same uplink bandwidth B, creating spatially dependent interference at the APs. Let $\mathcal{K} \triangleq \{1,\dots,K\}$ and $\mathcal{D} \triangleq \{1,\dots,D\}$ denote the sets of UEs and D2D transmitters, respectively.

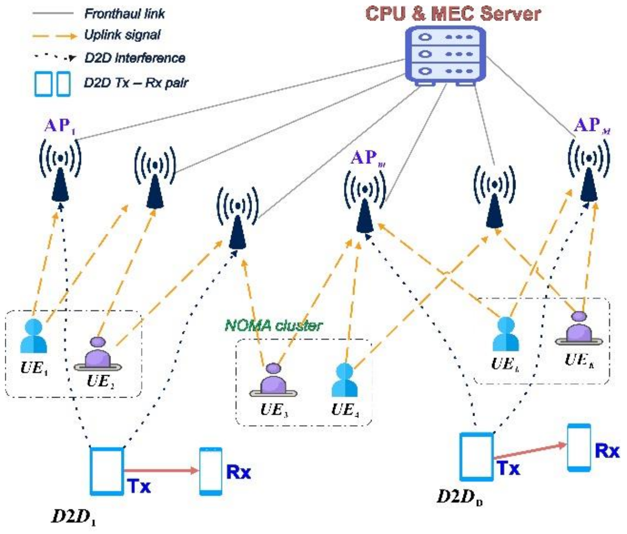

Figure 1. Architecture of the uplink single-carrier NOMA-CF-mMIMO network integrated with MEC under D2D underlay interference

The channel coefficient between UE k and AP m is modeled as $h_{mk} = \sqrt{\beta_{mk}}\tilde{h}_{mk}$, where β_{mk} represents large-scale fading, including path loss and shadowing, and $\tilde{h}_{mk} \sim \mathcal{CN}(0,1)$ denotes small-scale Rayleigh fading. Similarly, the interference channel from the d-th D2D transmitter to AP m is given by $g_{md} = \sqrt{\alpha_{md}}\tilde{g}_{md}$, with $\tilde{g}_{md} \sim \mathcal{CN}(0,1)$.

2.2. Uplink transmission and MRC combining

In the uplink, all K UEs and D D2D pairs transmit simultaneously over the same frequency band. The aggregate baseband signal received at AP m is given by

$$y_m = \sum_{k=1}^{K} \sqrt{P_k}\, h_{mk} s_k + \sum_{d=1}^{D} \sqrt{P_d}\, g_{md} x_d + n_m, \quad (1)$$

where, $P_k \in [0, P_{\max}]$ and P_d denote the transmit powers, in watts, of UE k and D2D transmitter d, respectively; s_k and x_d are normalized data symbols satisfying $\mathbb{E}[|s_k|^2] = 1$ and $\mathbb{E}[|x_d|^2] = 1$; and $n_m \sim \mathcal{CN}(0, \sigma^2)$ is the additive white Gaussian noise (AWGN) at AP m.

In this study, the transmit powers of the D2D transmitters are assumed to be fixed at given reference levels, corresponding to underlay D2D traffic that is scheduled and allocated spectrum independently of the cellular system, and thus they are not included in the set of optimization variables. Under this assumption, D2D interference appears to the CF-mMIMO system as an uncontrollable spatially structured background interference, while the transmission rates of the D2D links are preserved by the separate power allocation policy of the D2D layer.

The APs forward the received signal samples $\{y_m\}_{m=1}^{M}$ to the CPU, where maximum-ratio combining (MRC) is then applied to separate the signal of each UE. Let $\mathbf{h}_k = [h_{1k}, \dots, h_{Mk}]^T \in \mathbb{C}^M$ denote the aggregate channel vector of UE k and let $\mathbf{g}_d = [g_{1d}, \dots, g_{Md}]^T \in \mathbb{C}^M$ denote the aggregate D2D channel vector. The MRC combining vector is $\mathbf{v}_k = \mathbf{h}_k$. The effective channel gain after combining from UE j to the combiner of UE k is defined as:

$$H_{k,j} \triangleq |\mathbf{v}_k^H \mathbf{h}_k|^2\, \mathbf{1}\{j = k\} + |\mathbf{v}_k^H \mathbf{h}_j|^2\, \mathbf{1}\{j \neq k\}, \quad (2)$$

and the effective D2D interference gain is defined as

$$G_{k,d} \triangleq |\mathbf{v}_k^H \mathbf{g}_d|^2. \quad (3)$$

2.3. NOMA pairing and ECIR-based SIC model

To mitigate multi-user interference, the UEs are grouped into NOMA clusters of size two. Let $X_{i,j} \in \{0,1\}$ denote the binary pairing indicator variable: $X_{i,j} = 1$ if UE i and j form a cluster, and $X_{i,j} = 0$ otherwise. Let $\pi(k) \in \{1,2\}$ denote the SIC decoding order of UE k in its cluster, where $\pi(k) = 1$ indicates that UE k is decoded first, while the signal from its cluster partner is still treated as interference, and $\pi(k) = 2$ indicates that it is decoded second, after the signal of its partner has been removed.

In the presence of local D2D interference, conventional SIC ordering based on channel gain is no longer reliable. Therefore, we define the effective channel-to-interference ratio (ECIR) for UE k at the CPU as

$$\eta_k \triangleq \frac{H_{k,k}}{\sum_{d=1}^{D} P_d G_{k,d} + \sigma^2 \|\mathbf{v}_k\|^2}. \quad (4)$$

During SIC decoding at the CPU, UE k is affected by two types of interference: intra-cluster interference and inter-cluster interference. The signal-to-interference-plus-noise ratio (SINR) of UE k is given by

$$\gamma_k(\mathbf{P}, \mathbf{X}, \pi) = \frac{P_k H_{k,k}}{\sum_{j \in \mathcal{I}_k} P_j H_{k,j} + \underbrace{\sum_{d=1}^{D} P_d G_{k,d} + \sigma^2 \|\mathbf{v}_k\|^2}_{\triangleq \Theta_k}}, \quad (5)$$

where Θ_k denotes the aggregate D2D interference and the amplified noise component at the combiner, and $\mathcal{I}_k$ is the set of UEs causing intra-cluster interference to UE k. This set is determined by the pairing matrix X and the decoding-order vector π: a UE j belongs to the interference set of UE k if and only if UE j is in the same cluster as UE k and is decoded after UE k, meaning that the signal of UE j has not yet been removed at the time UE k is decoded. Therefore, the dependence of γ_k on $\mathbf{X}$ and π in (5) is reflected through the selection of this intra-cluster interference set.

The achievable uplink data rate of UE k is given by

$$R_k(\mathbf{P}, \mathbf{X}, \pi) = B\log_2(1 + \gamma_k(\mathbf{P}, \mathbf{X}, \pi)). \quad (6)$$

2.4. Partial computation offloading model

Each UE k intends to execute a computational task parameterized by the tuple $U_k \triangleq (L_k, C_k, T_k^{\max})$, where L_k is the input data size, in bits, C_k denotes the number of CPU cycles required per input bit, and $T_k^{\max}$ is the stringent latency deadline. We adopt a partial offloading architecture in which a variable fraction $\beta_k \in [0,1]$ of the task is processed locally, while the remaining fraction $(1-\beta_k)$ is offloaded through the CF-mMIMO uplink to the MEC server co-located with the CPU.

The local computation latency depends on the local CPU frequency f_k^{loc} and is given by

$$T_k^{\mathrm{loc}} = \frac{\beta_k L_k C_k}{f_k^{\mathrm{loc}}}. \quad (7)$$

The edge computation latency includes both the transmission latency and the execution latency at the MEC server. Under a fixed allocation policy that ensures fairness, we allocate the total MEC CPU capacity F_{CPU} equally among all K UEs, that is, the MEC CPU frequency assigned to UE k is $f_k^{\mathrm{MEC}} \triangleq F_{\mathrm{CPU}}/K$. The edge-side latency is then calculated as

$$T_k^{\mathrm{off}} = \frac{(1-\beta_k)L_k}{R_k} + \frac{(1-\beta_k)L_kC_k}{\frac{F_{\mathrm{CPU}}}{K}}. \tag{8}$$

To ensure that the task is completed before its deadline, and assuming that local and edge computations are performed in parallel, with the dominant component determining the latency, the overall latency constraint is given by

$$\max\{T_k^{\mathrm{loc}}, T_k^{\mathrm{off}}\} \le T_k^{\max} \tag{9}$$

3. Problem formulation

The objective of the problem is to maximize the system sum rate by jointly optimizing the NOMA pairing matrix $\mathbf{X} = [X_{i,j}]$, the SIC decoding-order vector $\pi = [\pi(1), \dots, \pi(K)]^T$, the local computation ratio vector $\beta = [\beta_1, \dots, \beta_K]^T$, and the transmit power vector $\mathbf{P} = [P_1, \dots, P_K]^T$. Substituting (7)–(9) into the system sum-rate expression, we obtain

$$\text{(P1):} \quad \max_{\mathbf{X},\pi,\beta,\mathbf{P}} \sum_{k=1}^{K} R_k(\mathbf{P}, \mathbf{X}, \pi) \tag{10a}$$

$$\text{s.t.} \quad \frac{\beta_k L_k C_k}{f_k^{\mathrm{loc}}} \le T_k^{\max},\ \forall k \in \mathcal{K}, \tag{10b}$$

$$\frac{(1-\beta_k)L_k}{R_k} + \frac{(1-\beta_k)L_kC_k}{f_k^{\mathrm{MEC}}} \le T_k^{\max},\ \forall k \in \mathcal{K}, \tag{10c}$$

$$0 \le P_k \le P_{\max},\ \forall k \in \mathcal{K}, \tag{10d}$$

$$0 \preccurlyeq \beta \preccurlyeq 1, \tag{10e}$$

$$X_{i,j} \in \{0,1\},\ \pi(k) \in \{1,2\},\ \forall i, j, k. \tag{10f}$$

Problem (P1) is a mixed-integer nonlinear programming (MINLP) problem. Finding its globally optimal solution is NP-hard due to the binary and permutation variables in (10f), the tight coupling between β_k and R_k in the fractional constraint (10c), and the non-convex multi-user interference components embedded in R_k through γ_k.

Task feasibility condition: For problem (P1) to have a feasible solution, the task parameters must satisfy the condition $T_k^{\max}(f_k^{\mathrm{loc}} + f_k^{\mathrm{MEC}}) \ge L_kC_k$ for all $k \in K$, i.e., the aggregate cooperative computing capability, consisting of the local CPU and the edge CPU allocated to the UE, must be sufficient to complete the task within the deadline. When this condition is violated, there exists no $\beta_k \in [0,1]$ and $R_k > 0$ that can simultaneously satisfy both (10b) and (10c). Throughout the following analysis, we assume that the task feasibility condition is guaranteed for all UEs.

4. Proposed alternating optimization algorithm

To handle the intractability of problem (P1), we decompose the original problem into three tractable subproblems and solve them iteratively within an alternating optimization (AO) framework.

4.1. Subproblem 1: ECIR-based NOMA pairing and SIC ordering

For SIC to operate effectively, UEs with clearly distinct effective channel conditions should be paired together. Using the proposed ECIR parameter η_k in (4), we sort all K UEs in descending order such that $\eta_{\sigma(1)} \ge \eta_{\sigma(2)} \ge \dots \ge \eta_{\sigma(K)}$, where $\sigma(\cdot)$ denotes the sorted index mapping.

A strong–weak greedy pairing algorithm is applied: the UE with the highest ECIR is paired with the UE with the lowest ECIR. Within each pair (i, j) with $\eta_i > \eta_j$, to ensure successful SIC and mitigate error propagation, UE i, i.e., the UE with the stronger ECIR, is decoded first while the signal of UE j remains present, namely $\pi(i) = 1$, whereas UE j is decoded last after the interference from UE i has been removed, namely $\pi(j) = 2$. This step determines $\mathbf{X}$ and π, and consequently identifies the interfering-user set $\mathcal{I}_k$.

4.2. Subproblem 2: Local computation allocation

Given the transmit power vector $\mathbf{P}$ and a fixed pairing structure, the rate R_k is determined. To reduce the transmission burden on the CF-mMIMO uplink congested by D2D interference, the system should maximize the local computation ratio β_k. From constraint (10b), the upper bound of β_k imposed by the local-processing latency limit is $T_k^{\max} f_k^{\mathrm{loc}}/(L_kC_k)$. Therefore, we choose

$$\beta_k^\star = \min\left(1, \frac{T_k^{\max} f_k^{\mathrm{loc}}}{L_kC_k}\right). \tag{11}$$

Optimality remark: The choice in (11) is not merely heuristic but is in fact optimal for problem (P1). Specifically, β_k does not appear in the objective function (10a) and only participates in the latency constraints (10b)–(10c); meanwhile, $R_k^{\mathrm{req}}(\beta_k)$ in (12) is monotonically decreasing with respect to β_k. Therefore, maximizing β_k is equivalent to minimizing the lower bound R_k^{req}, thereby maximally expanding the feasible region of the power allocation problem and ensuring that the achieved sum rate is no smaller than that of any alternative scheme with a lower β_k.

Given $\beta_k^\star$, constraint (10c) imposes the following minimum uplink data-rate requirement:

$$R_k^{\mathrm{req}} = \begin{cases} 0, & \text{if } \beta_k^\star = 1, \\ \dfrac{(1-\beta_k^\star)L_k}{T_k^{\max} - \frac{(1-\beta_k^\star)L_kC_k}{\frac{F_{\mathrm{CPU}}}{K}}}, & \text{otherwise}, \end{cases} \tag{12}$$

This serves as a lower-bound QoS requirement for the subsequent power control step. When $\beta_k^\star = 1$, the UE processes its entire task locally, and the uplink transmission rate is not constrained by latency, being limited only by $P_{\max}$.

4.3. Subproblem 3: SCA-based transmit power control

Given $\mathbf{X}$, π, $\beta^\star$, and R_k^{req} from the previous subproblems, problem (P1) reduces to the following transmit power control problem:

$$\max_{\mathbf{P}} \sum_{k=1}^{K} R_k(\mathbf{P})$$

$$\text{s.t.} \quad R_k(\mathbf{P}) \ge R_k^{\mathrm{req}}, \forall k \in \mathcal{K}, \tag{13}$$

$$0 \le P_k \le P_{\max}, \forall k \in \mathcal{K}. \tag{14}$$

The objective function is non-convex due to the interference components embedded in γ_k. We exploit the fact that $R_k(\mathbf{P})$ can be decomposed as the difference of two concave (DC) functions:

$$R_k(\mathbf{P}) = A_k(\mathbf{P}) - B_k(\mathbf{P}), \quad (15)$$

where

$$A_k(\mathbf{P}) = B\log_2\left(\sum_{j\in\{k\}\cup\mathcal{I}_k} P_j H_{k,j} + \Theta_k\right), \quad (16)$$

$$B_k(\mathbf{P}) = B\log_2\left(\sum_{j\in\mathcal{I}_k} P_j H_{k,j} + \Theta_k\right). \quad (17)$$

Both A_k and B_k are concave functions with respect to $\mathbf{P}$. To linearize the subtracted component B_k, we apply the first-order Taylor expansion at the feasible point $\mathbf{P}^{(t)}$ in the t-th iteration, yielding the following global linear upper bound:

$$B_k(\mathbf{P}) \leq \tilde{B}_k(\mathbf{P}) \triangleq B_k\left(\mathbf{P}^{(t)}\right) + \sum_{j\in\mathcal{I}_k} \nabla_j B_k(\mathbf{P}^{(t)})(P_j - P_j^{(t)}) \quad (18)$$

where the partial derivative is given by

$$\nabla_j B_k\left(\mathbf{P}^{(t)}\right) = \frac{B}{\ln 2} \cdot \frac{H_{k,j}}{\sum_{i\in\mathcal{I}_k} P_i^{(t)} H_{k,i} + \Theta_k}. \quad (19)$$

Replacing B_k with the linear surrogate function $\tilde{B}_k$ creates a global lower bound for the transmission rate, given by $\tilde{R}_k(\mathbf{P}) \triangleq A_k(\mathbf{P}) - \tilde{B}_k(\mathbf{P})$. Then, in the t-th iteration, problem (13)–(14) is approximated by the following convex problem:

$$(\mathrm{P}_{\mathrm{SCA}}^{(t)}): \quad \max_{\mathbf{P}} \sum_{k=1}^{K} \tilde{R}_k(\mathbf{P})$$

$$\text{s.t.} \quad \tilde{R}_k(\mathbf{P}) \geq R_k^{\mathrm{req}}, 0 \leq P_k \leq P_{\max}, \forall k, \quad (20)$$

This convex problem can be efficiently solved using standard interior-point solvers, such as the CVX toolbox.

4.4. Complexity analysis

In each AO iteration, the ECIR sorting step has a complexity of $\mathcal{O}(K\log K)$, the greedy pairing step has a complexity of $\mathcal{O}(K)$ and the local computation allocation step has a complexity of $\mathcal{O}(K)$. The convex subproblem in each SCA iteration has K optimization variables, corresponding to the power vector P and 3K constraints, consisting of K minimum-rate constraints and 2K power-bound constraints. Solving this problem using a standard primal–dual interior-point method, such as the SDPT3 solver integrated in CVX, has a complexity of $\mathcal{O}(K^{3.5}\log(1/\epsilon))$.

Therefore, the overall complexity of the algorithm is $\mathcal{O}(I_{\mathrm{conv}} \cdot K^{3.5}\log(1/\epsilon))$, where I_{conv} is the number of AO iterations required for convergence. The above bound corresponds to the cost of solving one convex problem; since each AO step includes an SCA loop, with each SCA iteration solving exactly one convex problem, the total computational cost of the algorithm is given by the product of the number of AO iterations, the average number of SCA iterations within each AO step, and the complexity of solving one convex problem as stated above.

5. Simulation results

5.1. Simulation setup

The performance of the proposed ECIR-NOMA technique is evaluated through Monte Carlo simulations. The UEs and D2D transmitters are uniformly distributed within a coverage area of 1 km × 1 km with a minimum AP–UE distance of $r_{\min} = 10$ m. Large-scale fading is modeled according to the 3GPP path-loss standard. The main simulation parameters are summarized in Table 1. Each data point in the result figures is averaged over 1,000 independent channel realizations; this number of samples is sufficiently large to reduce the 95% confidence interval of the average sum rate to below the performance gap among the compared techniques, thereby ensuring the statistical reliability of the results.

For comparison, we consider two benchmark systems:

● **No-NOMA**: The UEs are served individually, without clustering; hence, no intra-cluster interference exists.

● **Random-NOMA**: The UEs are paired into clusters according to a uniform random distribution.

Table 1. *Simulation parameter settings*

Parameter	Value
Coverage area	$1000 \times 1000\ m^2$
Number of APs (M)	32
Number of UEs (K)	10 (default)
Number of D2D pairs (D)	4 (default)
System bandwidth (B)	20 MHz
Noise power spectral density (N_0)	−174 dBm/Hz
Path-loss model	$35.3 + 37.6\log_{10}(d)$ dB
MEC CPU frequency (F_{CPU})	7.5 GHz
Local CPU frequency (f_k^{loc})	2.0 GHz
CPU cycles per bit (C_k)	1000 cycles/bit
Task size (L_k)	0.1 – 0.5 Mbit
Maximum task latency ($T_k^{\max}$)	200 ms

5.2. Performance evaluation

Sum rate versus D2D density: Figure 2 shows the system sum rate versus the number of D2D pairs, with the default configuration $M = 32, K = 10$. As D increases from 0 to 10, the sum rates of all techniques gradually decrease due to the increasing spatially heterogeneous D2D interference. The proposed ECIR-NOMA technique consistently outperforms both benchmark techniques over the entire range of D, achieving a sum-rate improvement of approximately 1.06%–1.16% over No-NOMA and 0.73%–0.78% over Random-NOMA. This relative performance gap slightly widens as D increases, confirming that ECIR-based SIC ordering, by explicitly accounting for the spatial distribution characteristics of D2D interference, can preserve the benefits of NOMA in interference-rich environments where conventional pairing methods are prone to failure.

The widening performance gap can be explained as follows: as the number of D2D pairs increases, the interference at the APs becomes stronger and more spatially uneven, causing SIC ordering criteria based solely on channel gain to increasingly select incorrect decoding orders. For example, a UE with a large channel gain may still suffer from severe D2D interference and thus may not be suitable for early decoding. In contrast, ECIR directly uses the actual signal-to-interference-plus-noise condition at the combiner, thereby maintaining a near-optimal ordering even under severe interference. This is why its relative gain increases precisely in the high-interference regime.

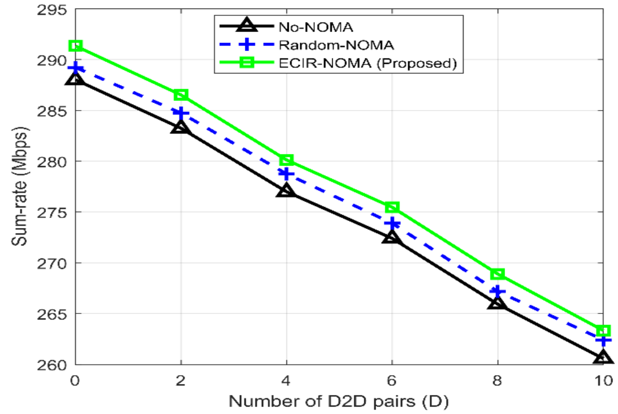

Figure 2. *System sum rate versus the number of D2D pairs*

Sum rate versus the number of UEs: Figure 3 shows the system sum rate versus the number of UEs, $K \in \{6, 8, \dots, 20\}$, with the number of D2D pairs fixed at $D = 4$. As K increases, the sum rate naturally increases due to the multi-user aggregation effect, although the SINR of each UE decreases as intra-system interference becomes denser. ECIR-NOMA maintains its advantage over the entire range of K, achieving a 1.00%–1.76% improvement over No-NOMA. Notably, the performance gap widens at moderate values of K, namely $K = 10$–20, where NOMA provides its most pronounced benefit: the number of UEs is sufficiently large for pairing to be meaningful, while cluster congestion has not yet become dominant.

In the large-UE regime, the gap narrows for two reasons: (i) intra-cluster interference and inter-user interference increase rapidly, saturating the separation gain of NOMA across all techniques; and (ii) the shared MEC computing resource, where ($f_k^{\mathrm{MEC}} \triangleq F_{\mathrm{CPU}}/K$ decreases with K) together with the latency budget, becomes a bottleneck, limiting the additional rate gain that optimized pairing can exploit.

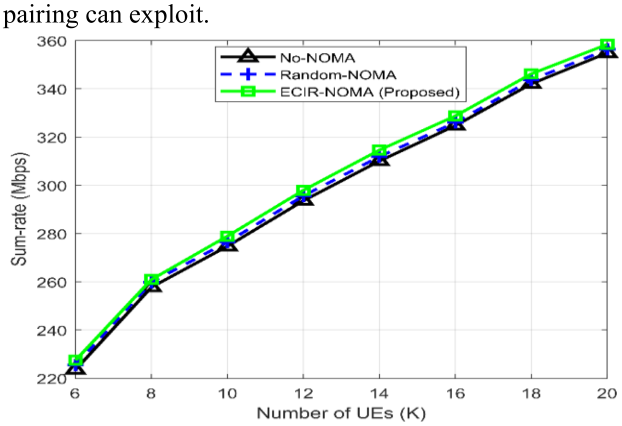

Figure 3. *Effect of the number of UEs on the system sum rate*

Algorithm convergence: Figure 4 shows the sum rate versus the number of iterations for a typical channel realization with $M = 32, K = 10, D = 4$. The proposed ECIR-NOMA algorithm converges rapidly and stably: it reaches approximately 95% of the steady-state sum-rate value after only the first 5 iterations and exceeds 99% by the 10th iteration, which is consistent with the theoretically guaranteed monotonic improvement property of SCA. The two benchmark systems, Random-NOMA and No-NOMA, also converge at a similar speed but settle at significantly lower sum-rate values, thereby reinforcing the validity of the ECIR-based pairing criterion. The fast convergence speed, requiring only 10–15 iterations, makes the proposed algorithm particularly suitable for MEC deployment scenarios with real-time constraints.

The observed monotonic improvement stems from the structure of the algorithm: each subproblem in the AO loop, including pairing, computation allocation, and power control, either improves or maintains the objective value, while the concave surrogate function used in SCA is always a lower bound of the true rate and is tightened after each iteration. Therefore, the sum rate does not decrease across iterations and converges to a stationary point.

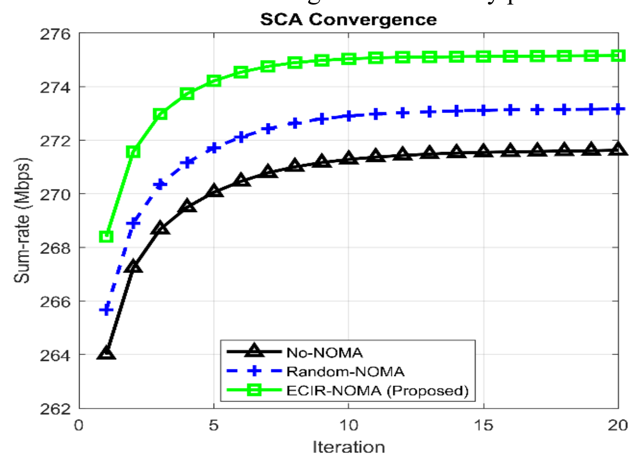

Figure 4. *Convergence behavior of the proposed AO-SCA algorithm, with M = 32, K = 10, and D = 4 D2D pairs*

6. Conclusion

This paper studied the problem of joint resource allocation and partial computation offloading in a single-carrier NOMA-CF-mMIMO MEC network under D2D spectrum-sharing interference. We showed that the performance of conventional channel-gain-based NOMA pairing degrades significantly in the presence of spatially dependent interference, and therefore proposed ECIR as a robust SIC ordering criterion that explicitly accounts for the D2D interference component at the output of the MRC combiner. A uniform MEC resource allocation policy among UEs was adopted to simplify the optimization process while preserving problem feasibility.

To solve the resulting non-convex MINLP problem, we developed a lightweight AO-SCA algorithm with a per-iteration complexity of $\mathcal{O}(K^{3.5}\log(1/\epsilon))$. Numerical simulation results confirmed that the proposed ECIR-NOMA technique reaches 99% of the steady-state sum-rate value within only 10–15 iterations and consistently outperforms the two benchmark systems, namely No-NOMA and Random-NOMA, while maintaining its advantage even under high-density D2D interference. Future work will extend this solution framework to multicarrier system configurations and environments with imperfect channel state information, where the tight coupling between the radio layer and the MEC layer becomes particularly important.

Acknowledgment: This research is funded by Vietnam National Foundation for Science and Technology Development (NAFOSTED) under grant number 102.04-2023.40.

REFERENCES

[1] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2322–2358, 2017.

[2] Q.-V. Pham, F. Fang, V. N. Ha, M. J. Piran, M. Le, L. B. Le, W.-J. Hwang, and Z. Ding, "A survey of multi-access edge computing in 5G and beyond: Fundamentals, technology integration, and state-of-the-art," *IEEE Access*, vol. 8, pp. 116974-117017, 2020.

[3] T. V. Thai, M. T. P. Le, H. V. Nguyen, O.-S. Shin, and M. G. Di Benedetto, "Next-generation MIMO empowered mobile edge computing: A comprehensive survey toward 6G systems," *Ad Hoc Networks*, vol. 182, p. 104095, 2026.

[4] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO versus small cells," *IEEE Trans. Wireless Commun*., vol. 16, no. 3, pp. 1834–1850, 2017.

[5] S. Mukherjee and J. Lee, "Edge computing-enabled cell-free massive MIMO systems," *IEEE Trans. Wireless Commun*., vol. 19, no. 4, pp. 2884-2899, 2020.

[6] G. Interdonato and S. Buzzi, "Joint optimization of uplink power and computational resources in mobile edge computing-enabled cell-free massive MIMO," *IEEE Trans. Commun*., vol. 72, no. 6, pp. 1804–1820, 2024.

[7] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, L. Hanzo, and P. Xiao, "On the performance of cell-free massive MIMO relying on adaptive NOMA/OMA mode-switching," *IEEE Trans. Commun*., vol. 68, no. 2, pp. 792–810, 2020.

[8] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas Commun*., vol. 35, no. 10, pp. 2181–2195, 2017.

[9] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1628–1656, 2017.

[10] X. Lin, R. W. Heath, and J. G. Andrews, "The interplay between massive MIMO and underlaid D2D networking," *IEEE Trans. Wireless Commun*., vol. 14, no. 6, pp. 3337–3351, 2015.

[11] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process*. Lett., vol. 21, no. 12, pp. 1501–1505, 2014.

[12] H. Xu, W. Xu, Z. Yang, J. Shi, and M. Chen, "Pilot reuse among D2D users in D2D underlaid massive MIMO systems," *IEEE Trans. Veh.* Technol., vol. 67, no. 1, pp. 467–482, 2018.

[13] C. You, K. Huang, H. Chae, and B.-H. Kim, "Energy-efficient resource allocation for mobile-edge computation offloading," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1397–1411, 2016.

[14] T. V. Thai, M. T. P. Le, H. V. Nguyen, and O.-S. Shin, "NOMA-aided cell-free massive MIMO with MEC: A trade-off between latency and energy consumption," in *IEEE Int. Conf. Consumer Electron.-Asia (ICCE-Asia)*, Danang, 2024.

[15] T. V. Thai, M. T. P. Le, H. V. Nguyen, and T. T. Huynh, "Latency-driven optimization in a multicarrier MEC-enabled cell-free massive MIMO network: A Greedy-NOMA pairing and successive convex approximation method," *The University of Danang - Journal of Science and Technology*, vol. 23, no. 5A, pp. 19–24, 2025.

[16] X.-T. Dang, M. T. P. Le, H. V. Nguyen, S. Chatzinotas, and O.-S. Shin, "Optimal user pairing approach for NOMA-based cell-free massive MIMO systems," *IEEE Trans. Veh. Technol*., vol. 72, no. 4, pp. 4751–4765, 2023.

[17] J. Zhao, Y. Liu, K. K. Chai, A. Nallanathan, Y. Chen, and Z. Han, "Spectrum allocation and power control for non-orthogonal multiple access in HetNets," *IEEE Trans. Wireless Commun*., vol. 16, no. 9, pp. 5825–5837, 2017.