LATENCY-DRIVEN OPTIMIZATION IN A MULTICARRIER MEC-ENABLED CELL-FREE MASSIVE MIMO NETWORK: A GREEDY-NOMA PAIRING AND SUCCESSIVE CONVEX APPROXIMATION METHOD

TỐI ƯU HÓA ĐỘ TRỄ TRONG HỆ THỐNG CELL-FREE MASSIVE MIMO ĐA SÓNG MANG TÍCH HỢP ĐIỆN TOÁN BIÊN DI ĐỘNG DỰA TRÊN PHƯƠNG PHÁP GHÉP CẶP NOMA THAM LAM VÀ XẤP XỈ LỒI LIÊN TIẾP

Tien V. Thai*, Mai T.P. Le, Hieu V. Nguyen, Tung T. Huynh

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

*Corresponding author: thaivantien@dut.udn.vn

(Received: March 19, 2025; Revised: April 15, 2025; Accepted: April 26, 2025)

DOI: 10.31130/ud-jst.2025.136

Abstract - This paper investigates the delay optimization problem in multi-carrier cell-free massive MIMO (CF-mMIMO) networks integrated with mobile edge computing (MEC), utilizing non-orthogonal multiple access (NOMA) techniques to exploit channel correlation among user groups. The authors propose a Greedy-NOMA pairing algorithm based on channel correlation to determine the optimal user pairs, thereby enhancing the efficiency of successive interference cancellation across multiple frequency channels. The objective is to minimize the maximum delay via a nonlinear optimization problem that jointly considers multi-channel power allocation and computational offloading ratios. The proposed solution consists of two steps: (i) user pairing using the Greedy-NOMA algorithm, and (ii) solving the resulting non-convex problem via successive convex approximation (SCA). Simulation results demonstrate a significant reduction in latency, achieving an improvement of up to 13.5% compared to conventional random NOMA and CF-mMIMO methods, thereby confirming the feasibility of the proposed approach for 5G/6G networks.

Key words - CF-mMIMO; mobile edge computing; successive convex approximation; NOMA pairing; latency optimization

Tóm tắt - Bài báo này nghiên cứu bài toán tối ưu hóa độ trễ trong mạng cell-free massive MIMO (CF-mMIMO) đa sóng mang tích hợp điện toán biên di động (MEC), ứng dụng kỹ thuật đa truy cập không trực giao (NOMA) để khai thác tương quan kênh truyền giữa các nhóm người dùng. Nhóm tác giả đề xuất thuật toán ghép cặp NOMA tham lam (Greedy-NOMA) dựa trên tương quan kênh, xác định cặp người dùng tối ưu nhằm nâng cao hiệu quả khử nhiễu liên tiếp trên nhiều kênh tần số. Mục tiêu là giảm thiểu độ trễ lớn nhất thông qua bài toán tối ưu phi tuyến, kết hợp phân bổ công suất đa kênh và tỷ lệ chuyển tải tính toán. Giải pháp gồm hai bước: (i) ghép cặp người dùng bằng Greedy-NOMA, và (ii) sử dụng phương pháp xấp xỉ lồi liên tiếp (SCA) để giải bài toán không lồi. Kết quả mô phỏng cho thấy, độ trễ giảm đáng kể, cải thiện đến 13,5% so với phương pháp NOMA ngẫu nhiên và CF-mMIMO thông thường, khẳng định tính khả thi cho mạng 5G/6G.

Từ khóa - Cell-free massive MIMO; điện toán biên di động; xấp xỉ lồi liên tiếp; ghép cặp NOMA; tối ưu hóa độ trễ

1. Introduction

In the context of fifth-generation (5G) and future sixth-generation (6G) mobile networks, the requirements for ultra-low latency, high throughput, and massive connectivity are becoming increasingly critical [1], [2]. Numerous real-time applications, such as autonomous vehicle control, augmented reality, and remote surgery, demand end-to-end latency of only a few milliseconds or even less.

To meet these stringent requirements, the cell-free massive MIMO (CF-mMIMO) architecture has been proposed as a promising solution to overcome the limitations of conventional cellular networks [3]. Instead of relying on fixed base stations, CF-mMIMO deploys a large number of distributed access points (APs) connected to a central processing unit. Recent studies have demonstrated that CF-mMIMO can significantly enhance spectral efficiency and reduce inter-cell interference compared to traditional cellular networks [4].

Alongside CF-mMIMO, mobile edge computing (MEC) technology has been developed to minimize processing latency by placing computational resources closer to end users [5]. According to [6], integrating MEC into CF-mMIMO systems enables user equipment (UE) to offload resource-intensive tasks to edge nodes, thereby substantially reducing response time and energy consumption. Recent works such as [7] and [8] have proposed optimal models for computational resource allocation in CF-mMIMO networks combined with MEC. However, these studies have not fully addressed the aspect of latency optimization.

To further enhance system performance, non-orthogonal multiple access (NOMA) technology has been proposed for integration into CF-mMIMO systems [9]. Unlike traditional orthogonal access methods, NOMA allows multiple users to share the same time-frequency resources through power domain multiplexing. The efficiency of NOMA largely depends on the user pairing algorithm, especially in large-scale networks [10], [11].

Recent studies have proposed various solutions to minimize latency in next-generation wireless systems. Bennis et al. [12] introduced resource allocation optimization methods for ultra-low latency networks, focusing on achieving high reliability. Although this solution provides an important foundation, it does not jointly consider both transmission and computation latency, resulting in limited performance in edge computing environments. Dinh et al. [13] proposed a priority-based MEC task scheduling approach combined with frequency scaling, which is effective in scenarios with non-uniform workloads but suffers significant performance degradation in dense multi-user environments, particularly as the number of UEs increases.

Mao et al. [14] investigated the trade-off between power and latency in multi-user MEC systems and developed a joint optimization method for communication and computation resources. However, this solution has high computational complexity and does not exploit the advantages of the CF-mMIMO architecture. In the context of NOMA, Dang and colleagues [10], [11] proposed optimal user pairing methods for CF-mMIMO systems supporting NOMA. These works focus on throughput maximization but do not address latency optimization when combined with MEC. Thai et al. [15] proposed a solution to balance latency and energy consumption in CF-mMIMO systems with NOMA and MEC support; however, this approach uses random NOMA pairing and considers only single-channel scenarios.

Although each of these technologies has demonstrated individual benefits, the effective integration of all three - CF-mMIMO, MEC, and NOMA - remains a significant research challenge. Existing works are still limited in simultaneously addressing three critical issues: (1) joint optimization of transmission and computation latency; (2) development of efficient NOMA user pairing algorithms, particularly for CF-mMIMO environments; and (3) scalable solutions for multi-channel systems with many users. Notably, prior studies such as [15], [16] mainly focus on energy optimization or latency-energy trade-offs, while transmission and computation aspects are often considered separately, lacking a unified model that accounts for both factors in multi-channel scenarios.

Therefore, this paper aims to optimize the maximum latency by employing a greedy pairing method for CF-mMIMO systems based on channel correlation, and by jointly optimizing transmit power and computational offloading ratios for MEC protocols. This approach is intended to improve performance compared to previously proposed methods, particularly in scenarios with a large number of users and stringent latency requirements.

Contributions of the paper

This paper aims to optimize latency in multicarrier CF-mMIMO systems integrated with MEC and NOMA, with the main contributions summarized as follows:

● Propose a Greedy-NOMA pairing algorithm based on the highest channel correlation, optimizing successive interference cancellation (SIC) efficiency on each subcarrier, thereby improving data rates and reducing transmission latency.

● Formulate a nonlinear optimization problem that jointly considers both multicarrier transmission latency and computation latency, subject to transmit power and offloading ratio constraints for the CF-mMIMO-MEC system.

● Combine the Greedy-NOMA algorithm for optimal user pairing with the successive convex approximation (SCA) method to solve the proposed non-convex resource allocation problem.

● Evaluate the effectiveness of the proposed approach through Monte Carlo simulations and comparisons with existing solutions. Numerical results demonstrate that our method significantly reduces latency while maintaining stability as the system scales.

Notations

Table 1 presents the mathematical symbols and operators used throughout the paper.

Table 1. *Symbols and operators*

Symbol	Description
$\mathbf{h}_i^H$	Hermitian conjugate
$\lVert\mathbf{h}\rVert^2$	Vector norm squared
$\mathcal{A}\setminus\mathcal{B}, \mathcal{A}\cup\mathcal{B}$	Set difference, set union
$\emptyset, \in$	Empty set, set membership
$\preceq, \arg\ max$	Vector inequality, argument of maximum
$\mathbb{C}^{M\times 1}, \mathcal{CN}(0,1)$	Complex vector space, complex normal distribution
$\epsilon, \triangleq$	SCA convergence threshold, "defined as" symbol

2. System Model

2.1. CF-mMIMO network model

The research considers an uplink CF-mMIMO system as illustrated in Figure 1. The system comprises M single-antenna APs distributed over a coverage area of size $D \times D$ serving K UE pairs over N frequency channel. Each pair consists of two UEs sharing the same channel resource via power-domain NOMA. The APs are connected to a Central Processing Unit (CPU) through high-capacity backhaul links. The CPU also acts as an MEC server, providing computing resources for offloaded tasks from the UEs. Each UE can transmit data over one or more channels and has two options for task processing: (i) local computation on the device, or (ii) offloading tasks to the MEC server at the CPU. The offloading decision depends on channel conditions, task size, and available computational resources.

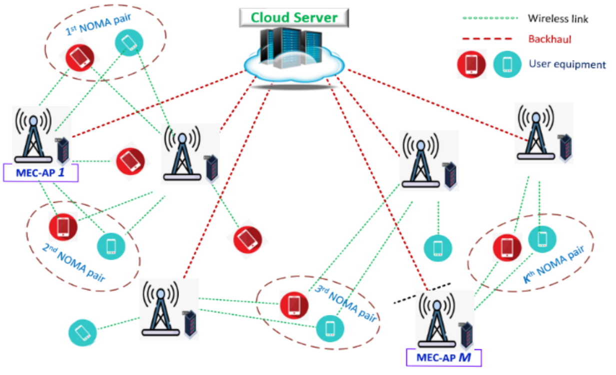

Figure 1. *CF-mMIMO system model combining MEC and NOMA*

In this system model, APs are assumed to only perform basic signal transmission/reception and maximum ratio combining (MRC). Other signal processing techniques, such as SIC, NOMA pairing, resource allocation, and computation processing, are executed centrally at the CPU. This architecture simplifies AP design, reduces deployment costs, and enables more efficient global optimization.

2.2. Channel model

The wireless channel between AP m and UE k on the n-th channel, denoted as $h_{m,n,k}$, is modeled as:

$$h_{m,n,k} = g_{m,n,k}\sqrt{\beta_{m,k}}, \tag{1}$$

where: $g_{m,n,k} \sim \mathcal{CN}(0,1)$ represents the small-scale Rayleigh fading, $\beta_{m,k}$ denotes the large-scale fading, defined as:

$$\beta_{m,k} = \frac{1}{d_{m,k}^{\alpha}}, \tag{2}$$

here, $d_{m,k}$ is the distance between AP m and UE k, α is the path-loss exponent.

It is assumed that the small-scale fading $g_{m,n,k}$ is independent across space (i.e., between different APs and UEs) and frequency (i.e., across different channels), while the large-scale fading $\beta_{m,k}$ remains constant for all channels for each AP-UE pair.

The overall channel vector from all APs to UE k on channel n is given by:

$$h_{n,k} = \left[h_{1,n,k}, h_{2,n,k}, \dots, h_{M,n,k}\right]^T \in \mathbb{C}^{M\times 1} \tag{3}$$

2.3. NOMA pairing and resource allocation

In uplink NOMA, paired UEs share the same time-frequency resources and are distinguished at the receiver by different power allocations. Specifically, a SIC receiver is used to remove the stronger UE's signal from the received mixture before decoding the weaker UE's signal. To enhance NOMA performance, instead of random pairing, the authors employ a greedy-NOMA pairing strategy based on channel characteristics.

Assume UE k_1 and k_2 are paired for SIC and share channel n, where the signal from UE k_1 is decoded before that of UE k_2. The allocated powers for these UEs are denoted as P_{n,k_1} and P_{n,k_2}, respectively. The received signal at AP m is:

$$\begin{aligned} y_{n,m} = {} & h_{m,n,k_1}\sqrt{P_{n,k_1}}x_{n,k_1} + h_{m,n,k_2}\sqrt{P_{n,k_2}}x_{n,k_2} \\ & + \textstyle\sum_{k'\neq k}^{K}(h_{m,n,k'_1}\sqrt{P_{n,k'_1}}x_{n,k'_1} + \\ & h_{m,n,k'_2}\sqrt{P_{n,k'_2}}x_{n,k'_2}) + z_n, \end{aligned} \tag{4}$$

where x_{n,k_1} and x_{n,k_2} are the transmitted signals of UEs k_1 and k_2 , z_n is additive white Gaussian noise (AWGN) with zero mean and variance σ_n^2.

At the CPU, the signal-to-interference-plus-noise ratio (SINR) for decoding UE k_1's signal is given by:

$$\mathrm{SINR}_{n,k_1} = \frac{P_{n,k_1}\left|\sum_{m=1}^{M} a_{m,n,k_1}h_{m,n,k_1}\right|^2}{\sum_{m=1}^{M}\left|a_{m,n,k_1}h_{m,n,k_2}\right|^2 P_{n,k_2} + \mathrm{IN}_1 + \sigma_n^2}, \tag{5}$$

where

$$\begin{aligned} \mathrm{IN}_1 \triangleq \textstyle\sum_{k'\neq k}^{K}\sum_{m=1}^{M} \Big(& \left|a_{m,n,k_1}h_{m,n,k'_1}\right|^2 P_{n,k'_1} + \\ & \left|a_{m,n,k_1}h_{m,n,k'_2}\right|^2 P_{n,k'_2} \Big) \end{aligned}$$

denotes interference from other UE pairs. The combining coefficients a_{m,n,k_x} are based on MRC.

After successfully decoding and subtracting UE k_1's signal using SIC, the SINR for UE k_2 is:

$$\mathrm{SINR}_{n,k_2} = \frac{P_{n,k_2}\left|\sum_{m=1}^{M} a_{m,n,k_2}h_{m,n,k_2}\right|^2}{\mathrm{IN}_2 + \sigma_n^2}, \tag{6}$$

where

$$\begin{aligned} \mathrm{IN}_2 \triangleq \sum_{k'\neq k}^{K}\sum_{m=1}^{M} \Big(& \left|a_{m,n,k_2}h_{m,n,k'_1}\right|^2 P_{n,k'_1} \\ & + \left|a_{m,n,k_2}h_{m,n,k'_2}\right|^2 P_{n,k'_2} \Big). \end{aligned}$$

2.4. Delay and data rate model

The total delay for UE k can be expressed as:

$$\begin{aligned} L_k &= max\{L_{\mathrm{trans},k_x} + L_{\mathrm{comp},k_x}\} \\ &= max\left\{\frac{\rho_{k_x}D_{k_x}}{R_{k_x}} + \frac{\rho_{k_x}D_{k_x}C_{k_x}}{f_{\mathrm{serv}}}\right\}, \end{aligned} \tag{7}$$

with $x \in \{1, 2\}$. Where, L_{trans,k_x} and L_{comp,k_x} are the transmission delay and computation delay, respectively. Queueing delay is considered negligible compared to transmission and computation delays and thus is omitted in (7). $\rho_{k_x} \in [0,1]$ is the offloading ratio, D_{k_x}(bit) is the task size, C_{k_x} is the number of CPU cycles required per bit, and f_{serv} is the processing frequency of the MEC server.

The achievable data rate for UE k_x on channel n is given by the Shannon formula:

$$R_{n,k_x} = B\log_2\left(1 + \mathrm{SINR}_{n,k_x}\right), \tag{8}$$

where B is the bandwidth of each channel.

The total data rate for UE k_x across all channels is:

$$R_{k_x} = \textstyle\sum_{n=1}^{N} R_{n,k_x} \tag{9}$$

3. Delay optimization problem description

Let the sets of power allocation variables and offloading ratios be denoted as $\mathbf{p} = \{P_{n,k_x}\}_{n\in\mathcal{N},k\in\mathcal{K},x\in\{1,2\}}$ and $\boldsymbol{\rho} = \{\rho_{n,k_x}\}_{n\in\mathcal{N},k\in\mathcal{K},x\in\{1,2\}}$, respectively. The delay optimization problem is formulated as follows:

$$\min_{\{\mathbf{p},\boldsymbol{\rho}\}} \max_{k=1,\dots,K} L_k \tag{10}$$

$$\text{s.t. } 0 \leq \sum_{n=1}^{N} P_{n,k_1} \leq P_{k_1}^{max}, \forall k, \tag{10a}$$

$$0 \leq \sum_{n=1}^{N} P_{n,k_2} \leq P_{k_2}^{max}, \forall k, \tag{10b}$$

$$0 \leq \rho \leq 1. \tag{10c}$$

Constraints (10a) and (10b) ensure that the transmit power of each UE does not exceed its maximum allowable power. Constraint (10c) guarantees that the offloading ratio remains within the valid range.

4. Proposed two-step algorithm for delay optimization

4.1. Greedy-NOMA pairing

First, this paper proposes a greedy-NOMA pairing algorithm to determine pairs of UEs that will use the SIC technique. In CF-mMIMO systems, APs often apply the MRC technique to leverage local computation capabilities. However, this can result in substantial interference between UEs with highly correlated channels.

To address this, the authors define two sets:

- $\mathcal{P}$: the set of paired UEs,
- $\mathcal{U}$: the set of unpaired UEs.

In each iteration i of the algorithm, the two UEs in set $\mathcal{U}$ with the highest channel correlation are paired and moved to set $\mathcal{P}$.

The greedy pairing algorithm is detailed in Algorithm 1 below.

Algorithm 1: Greedy-NOMA pairing

```
Initialize: Unpaired UE set $\mathcal{U} = \{1,2,\dots,2K\}$, Paired UE set $\mathcal{P} = \emptyset$
While $|\mathcal{U}| > 1$ do
    Calculate the channel correlation matrix $\mathbf{C}$ with elements:
        $C_{i,j} = \frac{|\mathbf{h}_i^H \mathbf{h}_j|^2}{\|\mathbf{h}_i\|^2 \|\mathbf{h}_j\|^2}, \forall i,j \in \mathcal{U}.$
    Find the pair $(i^*, j^*) = \arg \max_{i,j \in \mathcal{U}, i \neq j} C_{i,j}$
    Pair UE $i^*$ and $j^*$
    Update $\mathcal{U} = \mathcal{U} \setminus \{i^*, j^*\}, \mathcal{P} = \mathcal{P} \cup \{(i^*, j^*)\}$
end while
return the set of UE pairs $\mathcal{P}$
```

4.2. Delay optimization via SCA and computational complexity

After determining the NOMA pairs, the authors solve the optimization problem (10) using the SCA algorithm [13].

Regarding the computational complexity of this two-step algorithm, it is easy to see that the complexity of the Greedy-NOMA algorithm is $\mathcal{O}(K^2)$, since it requires calculating a $2K \times 2K$ channel correlation matrix and finding the largest element in this matrix. For the SCA algorithm, each SCA iteration needs to solve a convex problem of size K, which typically costs $\mathcal{O}\left((\mathrm{v}^2 \mathrm{c}^{2.5} + \mathrm{c}^{3.5}) \log\left(\frac{1}{\epsilon}\right)\right)$, where $v = 2K(N+1)$ and $c = 4K + 1$ are the numbers of variables and constraints, respectively, with $2K$ being the number of users, and ϵ is the convergence threshold. Substituting these values, the overall complexity per iteration is $\mathcal{O}(K^{4.5}(N+1)^2 \log(1/\epsilon))$.

The SCA algorithm guarantees convergence to a local optimum of the original non-convex problem. This is ensured by the fact that each iteration of the SCA method either improves or maintains the objective function value, thus ensuring monotonic convergence.

5. Performance analysis

5.1. Simulation setup and evaluation methodology

5.1.1. Simulation setup

In this section, the authors evaluate the performance of the proposed solution through Monte Carlo numerical simulations. The system parameters are configured based on practical deployments of 5G/6G networks, as detailed in Table 2. To ensure statistical reliability, each scenario is simulated from 1,000 to several thousand times with random user locations and channel realizations.

To assess the effectiveness of the proposed two-step algorithm (Greedy-NOMA + SCA), the authors compare it with two baseline schemes:

● **Random-NOMA**: UEs are randomly paired, after which SCA is applied to optimize power and transmission rates.

● **Conventional CF-mMIMO**: Each UE exclusively uses separate channel resources (NOMA is not applied), and power and transmission rates are optimized using SCA.

Table 2. *Simulation parameter settings*

Parameter	Value
Simulation area	1 km × 1 km
Number of APs (M)	64
Number of UEs ($2K$)	[10, 20, 30, 40, 50]
Number of channels (N)	[2, 4, 6]
Bandwidth	20 MHz
Noise power	−174 dBm/Hz

5.1.2. Evaluation method

To thoroughly analyze system performance, the authors use the cumulative distribution function (CDF) of delay. The CDF allows us to evaluate not only the average value but also the entire probability distribution of delay, providing a comprehensive view of the system's ability to meet various delay requirements.

The explicit mathematical expression of the CDF is defined as follows:

$$F_L(l) = \Pr(L \leq l) = \frac{1}{2K} \sum_{k=1}^{K} \sum_{x=1}^{2} \mathbf{1}(L_{k_x} \leq l), \quad (11)$$

where $F_L(l)$ is the cumulative distribution function of the delay L, l is the considered delay threshold, L_{k_x} is the delay of user k_x (with $x \in \{1,2\}$ for each NOMA pair), $\mathbf{1}(\cdot)$ is the indicator function, which equals 1 if the condition inside is true and 0 otherwise, $2K$ is the total number of users in the system.

From the Monte Carlo simulation, the empirical CDF is calculated as follows:

$$\hat{F}_L(l) = \frac{1}{N_{MC} \times 2K} \sum_{i=1}^{N_{MC}} \sum_{k=1}^{K} \sum_{x=1}^{2} \mathbf{1}(L_{k_x}^{(i)} \leq l), \quad (12)$$

where $\hat{F}_L(l)$ is the estimated CDF from the simulation, N_{MC} is the number of Monte Carlo simulation runs, and $L_{k_x}^{(i)}$ is the delay of user k_x in the i-th simulation run.

5.2. Convergence analysis of the SCA algorithm

Figure 2 illustrates the convergence speed of the SCA algorithm with different channel configurations $N = \{2, 4, 6\}$, 20 UEs, and various MEC processing frequency settings. The results show that the algorithm converges rapidly within approximately 4-5 iterations for all configurations. The number of channels N significantly affects both the starting point and the final convergence value, with increasing N from 2 to 6 reducing the initial delay from $1.8 \times 10^{-4}s$ to $1.56 \times 10^{-4}s$ at $f_{\text{serv}} = 5$ GHz. However, the improvement diminishes according to the "law of diminishing returns".

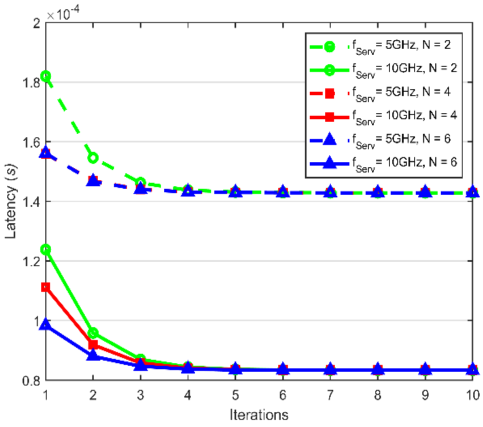

Figure 2. *Convergence analysis of the SCA algorithm*

In addition, the MEC processing frequency has a more substantial impact: increasing it from 5 GHz to 10 GHz reduces the converged delay from $1.43 \times 10^{-4}s$ to $0.83 \times 10^{-4}s$ (a 42% improvement) for $N = 2$. Notably, at $f_{\text{serv}} = 10$ GHz, the configurations with $N = 4$ and $N = 6$ achieve performance equivalent to $N = 2$, indicating that when computational resources are sufficiently large, increasing the number of channels does not provide significant benefits. The convergence curves decrease monotonically, with the most significant improvement observed in the first three iterations, demonstrating the effectiveness of the SCA method in quickly finding a locally optimal solution under various system conditions.

5.3. System delay evaluation

5.3.1. Delay distribution analysis

Figure 3 compares the delay of the three methods using CDF with the configuration $N = 2$, 20 UEs, and $f_{\text{serv}} = 5\ GHz$.

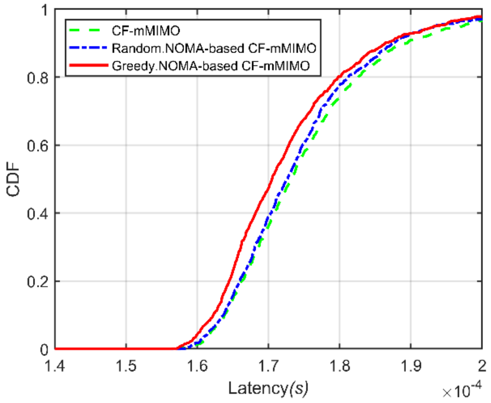

Figure 3. *Comparison of delay CDF between network architectures*

It is evident that the Greedy-NOMA method delivers superior performance compared to the other architectures. At the 80% user threshold, this solution achieves a delay of $1.78 \times 10^{-4}s$, which is lower than both Random-NOMA ($1.82 \times 10^{-4}s$) and conventional CF-mMIMO ($1.83 \times 10^{-4}s$). Notably, at a delay level of $1.7 \times 10^{-4}s$, the proposed method serves 70% of users, far surpassing Random-NOMA (50%) and conventional CF-mMIMO (45%). These results demonstrate the effectiveness of the Greedy-NOMA pairing algorithm in improving system delay.

5.3.2. Impact of the number of channels

Figure 4 analyzes the impact of the number of channels $N = \{2, 4, 6\}$ on the average delay with 20 UEs and $f_{\text{serv}} = 5$ GHz for the proposed Greedy-NOMA method.

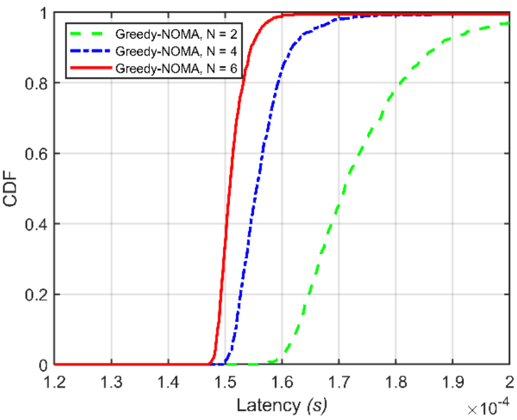

Figure 4. *Effect of number of channels $N = \{2, 4, 6\}$ on system delay*

The results show that the number of channels significantly affects system performance. Specifically, increasing N from 2 to 4, reduces the median delay from $1.8 \times 10^{-4}s$ to $1.62 \times 10^{-4}s$ (a 10% improvement). Further increasing to $N = 6$, the median delay continues to decrease to $1.52 \times 10^{-4}s$ (an additional 6% improvement). This "diminishing returns" trend suggests the need to balance performance and spectral resource efficiency.

Notably, at the delay threshold of $1.6 \times 10^{-4}s$, only about 30% of users achieve this with $N = 2$, while this figure rises to 80% with $N = 4$ and nearly 100% with $N = 6$. At the same time, the spread of the distribution also narrows as the number of channels increases, indicating higher fairness among users.

This improvement mechanism stems from UEs being able to distribute tasks over more channels, leveraging frequency diversity. These results are significant for system design: $N = 4$ may represent a good balance between performance and resource efficiency for low-latency applications in 5G/6G networks.

5.3.3. Scalability with respect to the number of users

Figure 5 provides a detailed analysis of the system's scalability as the number of users increases from 10 to 50, with $N = 2$ and various MEC processing frequency levels. At 10 users, the delays are $1.5 \times 10^{-4}s$, $1.2 \times 10^{-4}s$ and $1.1 \times 10^{-4}s$ corresponding to frequencies of 5 GHz, 7.5 GHz, and 10 GHz, respectively. As the number of users increases to 50, the delays rise significantly to $2.45 \times 10^{-4}s$, $2.1 \times 10^{-4}s$ and $1.85 \times 10^{-4}s$ for the respective frequency levels. Additionally, two significant trends are clearly observed. First, the delay increases with the number of users, as shown by the upward slope of all curves. Second, higher serving frequencies result in significantly lower delays at every user count, with the 10 GHz frequency curve consistently at the bottom and the 5 GHz curve always at the top.

These results clearly confirm the benefits of using higher frequencies to minimize system delay, especially as network scale increases. At 50 users, increasing the frequency from 5 GHz to 10 GHz reduces the delay by approximately 24.5%, which is a significant improvement for applications requiring low latency.

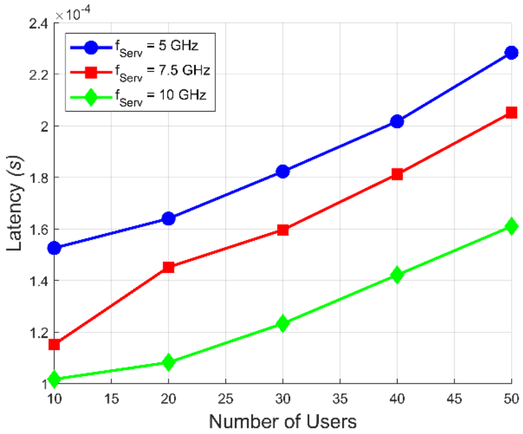

Figure 5. *Effect of number of users on system delay with number of channels* $N = 2$ *and* $f_{serv} = \{5, 7.5, 10\}GHz$

6. Conclusion

This paper has proposed a solution combining Greedy-NOMA pairing and SCA to optimize latency in CF-mMIMO networks integrated with MEC. Simulation results show that this method reduces latency by up to 13.5% compared to conventional CF-mMIMO and by 8.7% compared to random NOMA pairing. The algorithm converges quickly after 4–5 iterations and achieves the lowest latency of $0.83 \times 10^{-4}s$ at an MEC processing frequency of 10 GHz. When the number of users increases from 10 to 50, the latency only increases by 60%, demonstrating good scalability. Future work will focus on researching more optimal pairing methods or integrating machine learning for dynamic optimization, developing adaptive mechanisms for mobile environments, and expanding the model to simultaneously consider multiple QoS indicators, further enhancing the effectiveness of the solution in practical deployments of next-generation 5G/6G wireless networks.

Acknowledgment: This work was supported by The University of Danang - University of Science and Technology, code number of Project: T2024-02-40.

REFERENCES

[1] W. Saad, M. Bennis, and M. Chen, ‘A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems’, *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, 2020, doi: 10.1109/MNET.001.1900287.

[2] W. Jiang, B. Han, M. A. Habibi, and H. D. Schotten, ‘The Road Towards 6G: A Comprehensive Survey’, *IEEE Open J. Commun. Soc.*, vol. 2, pp. 334–366, 2021, doi: 10.1109/OJCOMS.2021.3057679.

[3] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, ‘Cell-Free Massive MIMO Versus Small Cells’, *IEEE Trans. Wirel. Commun.*, vol. 16, no. 3, pp. 1834–1850, 2017, doi: 10.1109/TWC.2017.2655515.

[4] J. Zhang, S. Chen, Y. Lin, J. Zheng, B. Ai, and L. Hanzo, ‘Cell-Free Massive MIMO: A New Next-Generation Paradigm’, *IEEE Access*, vol. 7, pp. 99878–99888, 2019, doi: 10.1109/ACCESS.2019.2930208.

[5] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, ‘A Survey on Mobile Edge Computing: The Communication Perspective’, *IEEE Commun. Surv. Tutor.*, vol. 19, no. 4, pp. 2322–2358, 2017, doi: 10.1109/COMST.2017.2745201.

[6] S. Mukherjee and J. Lee, ‘Edge Computing-Enabled Cell-Free Massive MIMO Systems’, *IEEE Trans. Wirel. Commun.*, vol. 19, no. 4, pp. 2884–2899, 2020, doi: 10.1109/TWC.2020.2968897.

[7] 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. 3, pp. 1804–1820, 2024, doi: 10.1109/TCOMM.2023.3289754.

[8] G. Femenias and F. Riera-Palou, ‘Mobile Edge Computing Aided Cell-Free Massive MIMO Networks’, *IEEE Trans. Mobile Comput.*, vol. 23, no. 2, pp. 1246–1261, 2024, doi: 10.1109/TMC.2023.3236543.

[9] A. Akbar, S. Jangsher, and F. A. Bhatti, ‘NOMA and 5G emerging technologies: A survey on issues and solution techniques’, *Comput. Netw.*, vol. 190, p. 107950, 2021, doi: https://doi.org/10.1016/j.comnet.2021.107950.

[10] X.-T. Dang, M. T. P. Le, H. V Nguyen, and O.-S. Shin, ‘Optimal User Pairing for NOMA-assisted Cell-Free Massive MIMO System’, in *2022 IEEE Int. Conf. Commun. Electron. (ICCE)*, 2022, pp. 7–12.

[11] 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, doi: 10.1109/TVT.2022.3222789.

[12] M. Bennis, M. Debbah, and H. V. Poor, ‘Ultrareliable and Low-Latency Wireless Communication: Tail, Risk, and Scale’, *Proceedings of the IEEE*, vol. 106, no. 10, pp. 1834–1853, 2018, doi: 10.1109/JPROC.2018.2867029.

[13] T. Q. Dinh, J. Tang, Q. D. La, and T. Q. S. Quek, ‘Offloading in Mobile Edge Computing: Task Allocation and Computational Frequency Scaling’, *IEEE Transactions on Communications*, vol. 65, no. 8, pp. 3571–3584, 2017, doi: 10.1109/TCOMM.2017.2699660.

[14] Y. Mao, J. Zhang, S. H. Song, and K. B. Letaief, ‘Power-Delay Tradeoff in Multi-User Mobile-Edge Computing Systems’, in *2016 IEEE Global Communications Conference (GLOBECOM)*, 2016, pp. 1–6. doi: 10.1109/GLOCOM.2016.7842160.

[15] 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 *2024 IEEE Int. Conf. Consum. Electron-Asia (ICCE-Asia)*, Danang: IEEE, Nov. 2024, pp. 1–5. doi: 10.1109/ICCE-Asia63397.2024.10773979.

[16] H. V Nguyen, M. T. P. Le, T. D. Ho, P. V. Tuan, and H. Nguyen-Le, ‘Joint Latency Minimization and Power Allocation for MEC-Enabled MU-MISO Networks’, in *2024 10th Int. Conf. Commun. Electron. (ICCE)*, 2024, pp. 753–757. doi: 10.1109/ICCE62051.2024.10634654.