IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 27 July 2023, accepted 24 August 2023, date of publication 4 September 2023, date of current version 7 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3311508

RESEARCH ARTICLE

Cooperation Method Between CPUs in Large-Scale Cell-Free Massive MIMO for User-Centric RAN

AKIO IKAMI, NAOKI AIHARA, YU TSUKAMOTO, TAKAHIDE MURAKAMI, AND HIROYUKI SHINBO
KDDI Research, Inc., Saitama 356-8502, Japan

Corresponding author: Akio Ikami (ak-ikami@kddi.com)

This work was supported by the National Institute of Information and Communications Technology (NICT), Japan, under Grant 00401.

ABSTRACT We have been studying a user-centric radio access network (RAN) for the realization of uniform radio quality "anywhere anytime" with Cell-free massive MIMO (CF-mMIMO) technology. In user-centric RAN, the central processing unit (CPU) that processes CF-mMIMO signals is assumed to be deployed in multiple sites to address the scalability problem for large-scale CF-mMIMO. However, the distributed deployment of CPUs results in radio quality degradation due to interference between UEs connected to CPUs at different sites. To address this problem, multiple CPU cooperation methods between CPUs at different sites are being studied. However, for cooperation, conventional methods require the exchange of radio signals and channel state information between CPUs, which significantly increases the transmission load on the backhaul connecting the sites. To resolve this issue, we propose an inter-site CPU cooperation method that maintains high radio quality while reducing the amount of data transmitted between sites to suppress inter-site interference. The proposed method is realized by deploying a channel estimation processing function for inter-site interference at each site and suppressing inter-site interference independently. Furthermore, we introduce optimization management that adjusts the degree of cooperation among CPUs based on the proposed method according to the required radio quality and computation and transmission resources in the area. We evaluate the proposed method by computational simulation. We show that the proposed method can reduce the transmission load by 53% with the same area throughput compared to the existing CPU cooperation schemes.

INDEX TERMS Cell-free massive MIMO, user-centric RAN, RAN management, 6G.

I. INTRODUCTION

Various consortiums and standardization task groups have been actively studying use cases for the 6th generation mobile communication system (6G), which is expected to be commercially available around 2030 [1], [2]. According to these studies, one of the common use cases described is the coexistence of mobile robots and humans, an arrangement that will contribute to addressing the issue of labor shortages. Safety is the most critical factor for the coexistence of mobile robots and humans. For safety, constant monitoring and operation from the cloud via 6G are needed wherever the robot is. Therefore, uniform and high radio quality, anytime and anywhere, is needed in 6G.

However, the 5th Generation (5G) systems have cell-edge issues due to increasing path loss and inter-cell interference, leading to the degradation of radio quality at cell-edge areas. Cell-free massive MIMO (CF-mMIMO) has attracted attention as a promising technology that can solve the cell-edge problem [3]. CF-mMIMO involves deploying access points (APs) around user equipment (UE), with inter-AP cooperation for transmitting and receiving signals. A central processing unit (CPU) performs concentrated signal processing to/from the APs. The CPU can suppress inter-cell interference and address cell-edge issues through coordinated signal processing. The initial CF-mMIMO proposal connected all

The associate editor coordinating the review of this manuscript and approving it for publication was Xujie Li.

APs to a single CPU. This resulted in scalability issues with computational load in the CPU for processing radio signals from all APs and transmission load between the CPU and APs. To address the scalability problem, recent studies have involved the use of a distributed CPU architecture and an optimized method for selecting the access point (AP) for each user [4], [5], [6], [7], [8], [9], [10], [11], [12]. These methods are based on several factors including the received signal power, channel state, and user mobility.

We have been studying a user-centric radio access network (RAN) architecture that aims to achieve uniform radio quality using CF-mMIMO. Our goal is to deploy this architecture in urban areas with the help of a mobile network operator (MNO), as outlined in [13]. The user-centric RAN concept involves creating a logical network for each user on a physical infrastructure using virtualized RAN (vRAN) technology [14]. This is achieved by placing virtualized base station functions, specifically radio signal processing, called virtualized CPU (vCPU), and selecting a cluster of access points (AP clusters) to serve each user. The vCPU and AP cluster are deployed and selected for each user by a user-centric RAN intelligent controller (uRIC), which is responsible for managing and controlling the RAN. By optimizing the logical network consisting of the vCPU and AP cluster, user-centric RAN facilitates the efficient use of transmission link resources between multiple sites deployed by vCPUs and provides computational resources at each site along with high radio quality for users.

There is the problem of degradation in radio quality due to inter-site interference in a distributed deployment of CPUs required for large-scale CF-mMIMO in user-centric RAN. This is due to the difficulty of coherent signal processing for interference suppression between UEs connecting to CPUs at different sites. Several methods [7], [8], [9], [10], [11] have been studied to solve this problem by enabling cooperation between site-to-site CPUs. When these existing methods are applied, exchanging radio signals and channel state information is necessary for multiple CPU cooperation. This exchange increases the load on transport links, leading to the same scalability problem that was the case with deployment of a single CPU.

To address this problem, we propose an inter-site CPU cooperation method that maintains high radio quality while reducing the volume of data transmitted between sites, which is necessary to suppress inter-site interference. The proposed method achieves its goals by sharing a list of APs whose inter-site interference exceeds the threshold and the pilot assignment information allocated to the interference source UEs, instead of using radio signals as in the existing CPU cooperation method. This shared information is used for channel estimation and interference suppression processes to reduce inter-site interference on a site-by-site basis. By performing independent channel estimation and interference suppression processes, the proposed method achieves uniform radio quality and reduces the transmission load between sites. Compared to the existing CPU cooperation method, the proposed method incurs a lower transmission load because it shares information instead of radio signals. Furthermore, we introduce optimization management that adjusts the degree of cooperation among CPUs based on the proposed method to generate the minimum transmission line load and impose a lower computational load for the required radio quality in the area, assuming a user-centric RAN that simulates the actual physical structure.

The remainder of this paper is structured as follows. Section II provides an overview of the user-centric RAN architecture proposed by the authors and highlights the challenges posed by inter-site interference in this architecture. In Section III, we present an inter-site CPU cooperation method that suppresses inter-site interference while minimizing the amount of data transmitted between sites. We then formulate an optimization problem that adjusts the level of CPU cooperation based on the proposed method to minimize transmission load. In Section IV, we evaluate the effectiveness of the proposed method in terms of radio quality, computational load, and transmission load. Finally, in Section V, we conclude the paper.

II. ARCHITECTURE OF USER-CENTRIC RAN

In this section, we describe the proposed user-centric RAN architecture and multiple CPU cooperation in large-scale CF-mMIMO and its problems.

A. STRUCTURE

Fig. 1 shows the architecture of the user-centric RAN designed to achieve uniform radio quality using CF-mMIMO. The concept of user-centric RAN involves creating a logical network for each user on a physical infrastructure using virtualized base station functions, such as vCPU. The physical topology assumed in this study is a double-star type topology, which is typical of optical access networks [15], [16]. The L APs placed in a given area are connected to J edge sites via optical fiber, and the radio signals from the APs are aggregated at the edge site. Each edge site is connected to a central site that aggregates all traffic in the area, and the traffic from each edge site is then aggregated at the central site via a backhaul (BH). The index function, SiteAP(l), indicates the edge site where AP l is located. Commodity servers with CPU functions are placed at each edge site to instantiate the vCPU per user, which processes signals from the APs.

In user-centric RAN, a per-user logical network is created by allocating physical resources to manage radio quality, computation load, and transmission load. The logical network consists of an AP cluster and a vCPU and is managed by a uRIC responsible for controlling the RAN. The AP cluster is the set of APs per user that send and receive radio signals. AP clustering reduces the computational load for signal processing by limiting the number of APs for each user. To address inter-site interference, user-centric RAN adopts the approach of forming AP clusters across sites [7], [8], [9], [10], [11]. In Fig. 2 (a), we show inter-site interference between UEs deploying vCPUs at different sites when

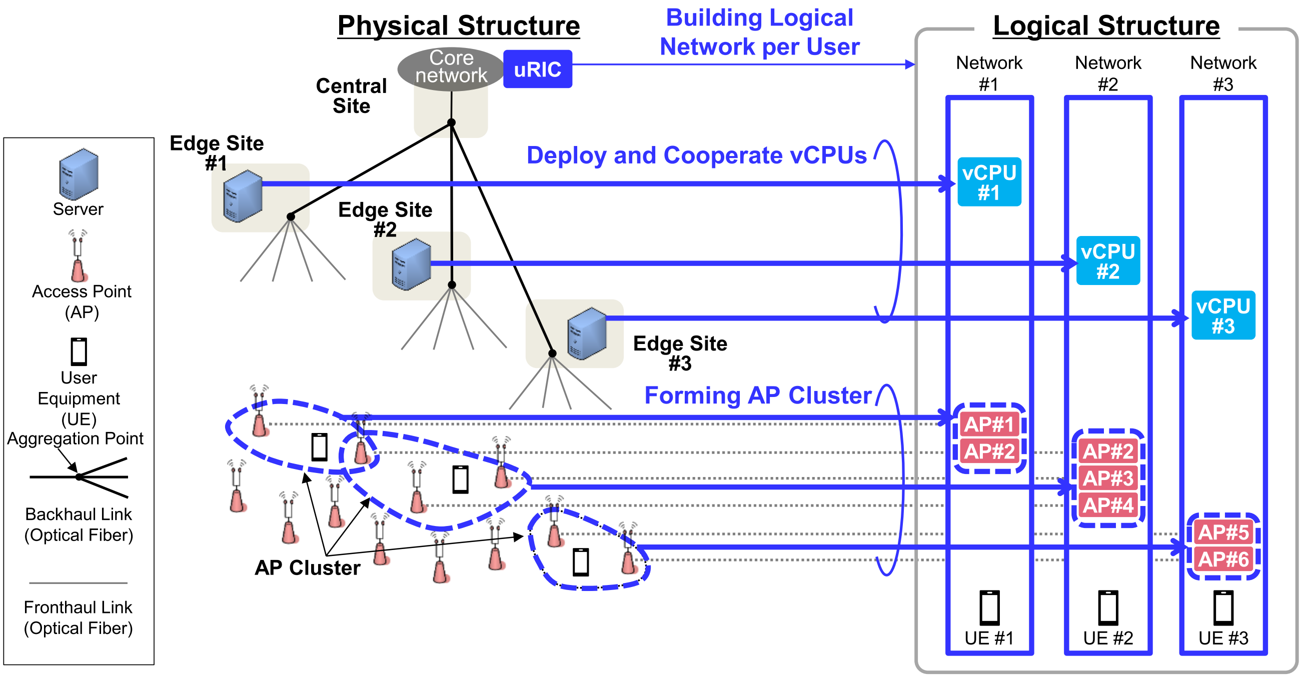


FIGURE 1. Architecture of user-centric RAN with three UEs. To balance radio quality for each user and computational and transmission load, a logical network per user is built by allocating physical resources in RAN. The logical network includes an AP cluster and a vCPU and is managed by a uRIC responsible for controlling the RAN.

vCPUs are distributed and there is no CPU cooperation between the different sites. Suppression of inter-site interference is a major challenge in user-centric RAN, which aims to maintain uniform radio quality everywhere since vCPUs are distributed across many sites, and there are many inter-site boundaries. As shown in Fig. 2 (b), this approach aggregates the radio signals from the APs that compose the AP cluster across sites where the vCPU is located for each UE to one site. Since the vCPU can form the weights for signal processing using the signals from APs connected to different sites, inter-site interference can be suppressed. The transport link in BH carries two types of data assuming that AP clusters are formed across sites. There are two types of data transfer in the user-centric RAN architecture. The first is radio signals transferred between APs and the vCPU, while the second is IP data containing user data transferred between the vCPU and the core network. Radio signals impose a greater load on the transport link than IP data.

B. MATHEMATICAL FORMULATION

We describe the mathematical formulation of the AP cluster, user throughput, computational load, and transmission load in BH involved in user-centric RAN. We consider K single-antenna UEs, and each UE k connects to AP l with the highest power by measuring the periodically broadcast synchronization signals. An initial connection between the AP and the UE is based on the method proposed in [8]. The initial connection scheme does not affect the inter-site interference that is the issue of this paper. Therefore, it is outside the scope of the proposed cooperation method between CPUs. We assume that the vCPU of UE k is placed in $\mathrm{Site}_{\mathrm{AP}}(l)$ connecting AP l.

AP clusters are defined according to the method proposed in [8]. The AP index belonging to an AP cluster $\boldsymbol{D}_k$ is defined as the following L-dimensional square matrix,

$$\boldsymbol{D}_k = \begin{bmatrix} D_{k1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & D_{kL} \end{bmatrix}, \tag{1}$$

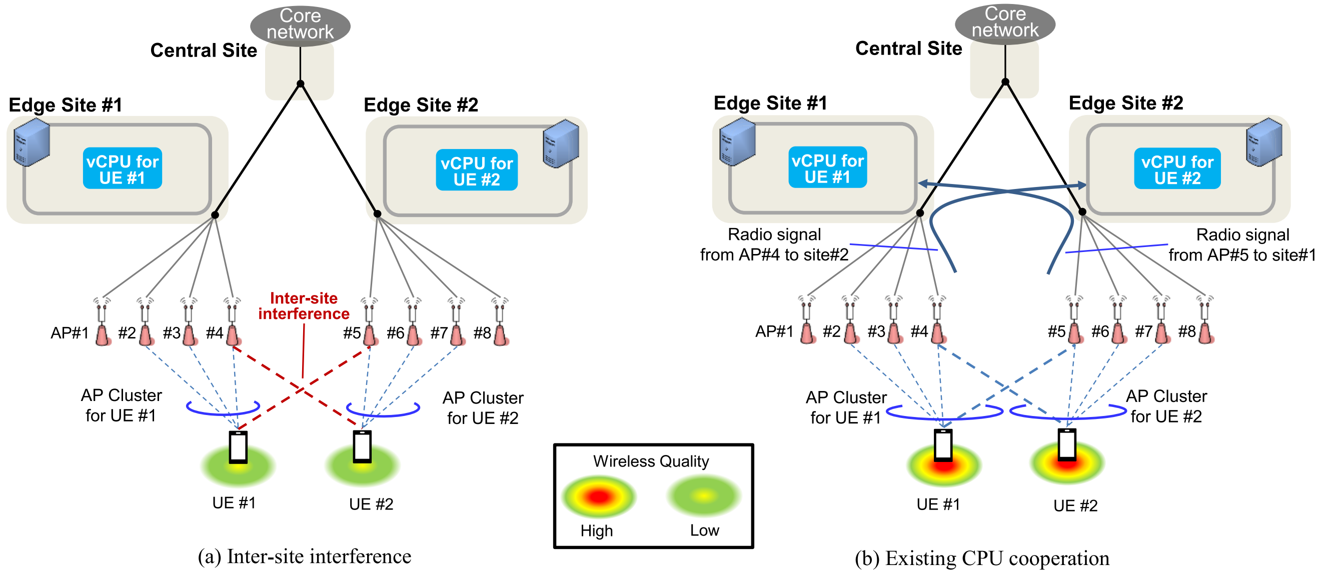
where D_{kl} is defined as

$$D_{kl} = \begin{cases} 1 & \text{if AP } l \text{ serves UE } k, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Let $\mathcal{M}_k$ be the set of APs where D_{kl} =1, which forms the AP cluster for UE k. The signal-to-interference and noise ratio (SINR) of the uplink for UE k is defined as follows [4]:

$$\mathrm{SINR}_k = \frac{p_k \left| v_k^H \boldsymbol{D}_k \hat{\boldsymbol{h}}_k \right|^2}{\sum_{i=1, i \neq k}^{K} p_i \left| \boldsymbol{v}_k^H \boldsymbol{D}_k \hat{\boldsymbol{h}}_i \right|^2 + \boldsymbol{v}_k^H \boldsymbol{Z}_k \boldsymbol{v}_k}, \tag{3}$$

where $\boldsymbol{Z}_k = \boldsymbol{D}_k \left(\sum_{i=1}^{K} p_i \boldsymbol{C}_i + \sigma^2 \boldsymbol{I}_L \right) \boldsymbol{D}_k$, p_k is the power of the uplink signal of UE k, and $\hat{\boldsymbol{h}}_i$ is the estimated channel coefficient, respectively. The channel coefficients are estimated using the standard minimum mean square error (MMSE) of Gaussian random variables [17]. We use the pilot assignment method [8] for large-scale CF-mMIMO. To minimize the error of channel estimation, we fix the value of the pilot power such that it is equal to the maximum uplink transmit power. $\boldsymbol{C}_i$ indicates a matrix of the channel estimation error for UE i. It is obtained from the difference



(a) Inter-site interference

(b) Existing CPU cooperation

FIGURE 2. Inter-site interference and existing CPU cooperation approach.

between the spatial channel correlation matrix estimated with MMSE and a real one. σ^2 is the power of thermal noise, and $\boldsymbol{I}_L$ is the L-dimensional identity matrix. $\boldsymbol{v}_k$ is the combining vector for UE k, which is obtained from the following equation assuming that the uplink signal received by APs is demodulated with partial-minimum mean squared error (P-MMSE) [18], [19],

$$\boldsymbol{v}_k = p_k \left(\sum\nolimits_{i\in\mathcal{P}_k} p_i \boldsymbol{D}_k \hat{\boldsymbol{h}}_i \hat{\boldsymbol{h}}_i^H \boldsymbol{D}_k + \boldsymbol{Z}_k\right)^{\dagger} \boldsymbol{D}_k \boldsymbol{h}_k. \tag{4}$$

Here, $\mathcal{P}_k$ is the set of UEs where the AP cluster for the UE is formed with at least one AP as used in the AP cluster for UE k. It is expressed as $\mathcal{P}_k = \{i : \boldsymbol{D}_k \boldsymbol{D}_i \neq \boldsymbol{O}_L\}$, where $\boldsymbol{O}_L$ is the L-dimensional zero matrix. The uplink user throughput TP_k for UE k is calculated with SINR as

$$TP_k = W_{\mathrm{RF}} \log_2 (1 + \mathrm{SINR}_k). \tag{5}$$

Here, W_{RF} is the total bandwidth of the wireless link.

The load on the transport links in the BH, R_{BH}, is defined as the sum of the radio signals and IP data transferred between the central site and edge sites. It can be obtained from:

$$R_{\mathrm{BH}} = \sum\nolimits_{j=1}^{J} (R_{\mathrm{RS}}(j) + R_{\mathrm{IP}}(j)). \tag{6}$$

$R_{\mathrm{RS}}(j)$ represents the radio signal originating from the APs forwarded from site j to the other sites when the AP cluster crosses between sites. It is calculated by

$$R_{\mathrm{RS}}(j) = \varepsilon_{\mathrm{RF}} R_{\mathrm{AP}} n_{(\mathrm{AP},j)}, \tag{7}$$

where R_{AP} is the transmission load per one AP, defined by the bandwidth allocation and quantization bit rate of the radio signal [20]. $n_{(\mathrm{AP},j)}$ is the number of APs connected to site j that are included in the AP cluster of the UE whose vCPU is located at a site other than site j. In other words, $n_{(\mathrm{AP},j)}$ indicates the number of APs that aggregate radio signals to the vCPU at other sites, as the AP cluster crosses between sites. $R_{\mathrm{IP}}(j)$ is the IP signal originating from vCPUs serving area j, and it is calculated by

$$R_{\mathrm{IP}}(j) = \sum\nolimits_{k\in\mathcal{I}_j} TP_k, \tag{8}$$

where $\mathcal{I}_j$ is the set of UEs which are served by the vCPU on site j. The overhead of the IP signal, e.g., headers, is ignored for simplicity.

The total computational load required for the signal processing at each site is defined as the computational load C_{comp}, which is calculated using the following equation

$$C_{\mathrm{comp}} = \sum\nolimits_{k\in K} \{C_{\mathrm{est}}(k) + C_{\mathrm{decode}}(k)\} + \sum\nolimits_{j\in J} c_{\mathrm{const}}, \tag{9}$$

where C_{est} is the computational load required for the channel estimation of the UE, and C_{decode} is the computational load required for the signal processing of UE k. These are expressed as follows

$$C_{\mathrm{est}}(k) = \left(N\tau_p + N^2\right) |\mathcal{M}_k||\mathcal{P}_k|, \tag{10}$$

$$C_{\mathrm{decode}}(k) = \frac{(N|\mathcal{M}_k|)^2 + N|\mathcal{M}_k|}{2}|\mathcal{P}_k| + (N|\mathcal{M}_k|)^2 + \frac{(N|\mathcal{M}_k|)^3 - N|\mathcal{M}_k|}{3}. \tag{11}$$

Here, N is the number of antennas deployed in the AP, and τ_p is the number of pilot sequences. In addition, C_{const} is a fixed value that indicates the computational load required by the OS and other basic processes, which is empirically obtained by [21].

C. PROBLEMS OF INTER-SITE CPU COOPERATION

Recently, various methods have been proposed for interference management in CF-mMIMO [22]. Most approaches consider computational scalability and manage the selection of AP clusters for each user to control the degree of interference suppression [4], [5], [6], [7], [8], [9], [10], [11], [12], [23], [24], [25], [26], [27]. Considering mobility, there are rule-based approaches for identifying on AP clusters with low computational complexity [6], [8], [23]. In [8], uniformity is achieved by selecting APs with high reference signal power values as AP clusters within a specific range. There are also optimization approaches [9], [10], [11], [24] where the system throughput, power consumption, and fairness index of the entire area are used as objective functions. These have the problem of being computationally expensive when the number of UEs and APs increases. These optimization approaches can form the AP clusters with the most efficient interference suppression on a small scale. Other AP cluster selection methods based on game theory [25] and machine learning-based AP cluster selection methods have also been proposed [26], [27]. This paper focuses on distributed CPU deployments for a large-scale cell-free implementation. Assuming distributed CPU deployment, interference between UEs at different sites, i.e., inter-site interference, occurs. In order to control this interference, inter-site CPU cooperation is needed. In [7], signaling between CPUs and APs is proposed to form AP clusters in a system model in which CPUs are distributed. In [11], a method is proposed to optimize the overall downlink transmit power and share the power allocation information among the CPUs at each site. However, there is no mention of the uplink in this paper. In [8], the authors demonstrated quantitatively for the first time that the transmission line load and computational load are scalable in large-scale CF-mMIMO, with distributed CPUs performing independent signal processing and forming AP clusters for each user. In [9], the authors integrate [8] into the architecture of O-RAN [28] and show there is a trade-off between transmission load and radio quality between distributed CPUs. In [10], the authors proposed an algorithm to calculate the formation of AP clusters across sites so that the sum of SINRs of all UEs is maximized. As another approach, [12] proposed to optimize the overall allocation of frequency resources to each CPU and share the allocation information among CPUs in each site. However, the spatial multiplicity is inferior to [7], [8], [9], [10], and [11] due to the frequency separation between sites.

To suppress interference between UEs at different sites, existing CPU cooperation methods aggregate user radio signals via BH at one site. However, this approach requires duplicating the radio signals $R_{\mathrm{RS}}(j)$ for other sites at each site j and transferring the radio signals via BH. Since the data volume of radio signal $R_{\mathrm{RS}}(j)$ is significantly larger than that of IP signal $R_{\mathrm{IP}}(j)$, the transmission load on the BH will increase significantly according to equation (6). In other words, there is a trade-off problem: the more APs connecting to different sites are added to the AP cluster to suppress inter-site interference, the more the transmission load in the BH increases significantly. To address this trade-off problem, it is necessary to reduce the amount of inter-site transmission data required to suppress inter-site interference while maintaining high radio quality. Thus, RAN management is required to generate the minimum transmission line load and impose a lower computational load for the required radio quality in the area. In addition, it is necessary to evaluate and manage the radio quality, the transmission load in the BH, and the computational load in actual RAN topology between sites since the transmission load depends on the RAN topology.

III. PROPOSED COOPERATION METHOD BETWEEN CPUs

A. OVERVIEW OF PROPOSED METHOD

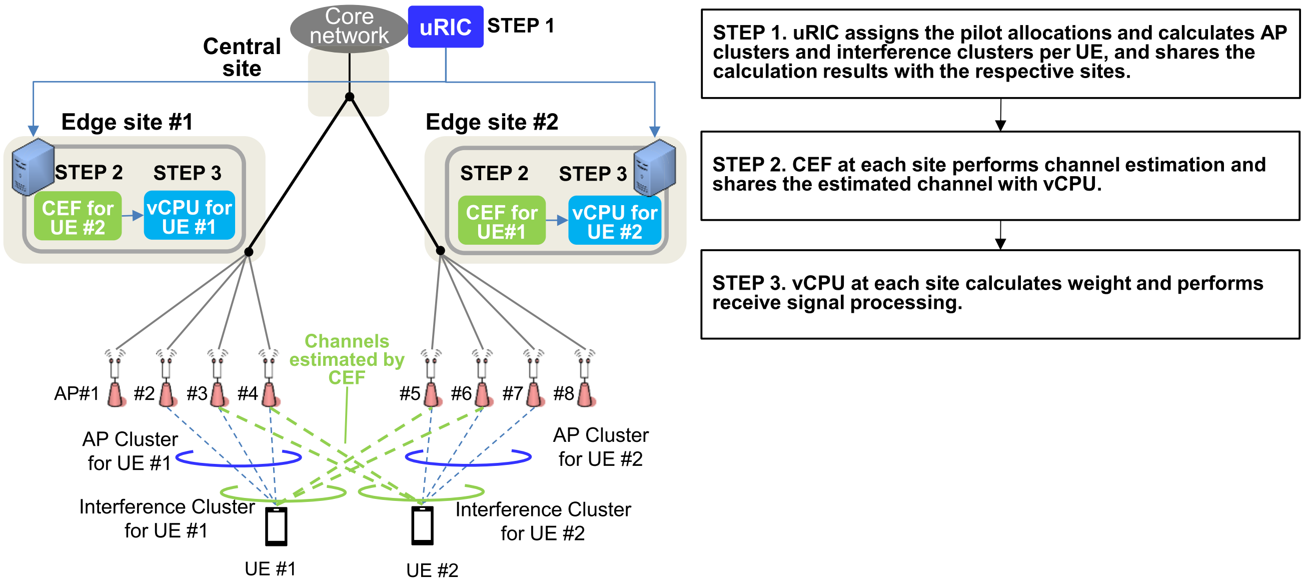
This section proposes an inter-site CPU cooperation scheme that maintains high radio quality while reducing the inter-site transmission data volume required to suppress inter-site interference. One of the approaches to suppress inter-site interference is coordinated beamforming [29], [30]. This scheme suppresses interference by sharing pilot allocation and channel information between neighboring cells. The proposed scheme is based on this idea. Fig. 3. shows the flow of the proposed method and provides an example for the case of two sites and two UEs. The proposed method is achieved by sharing a list of APs whose inter-site interference exceeds the threshold and the pilot assignment information allocated to the interference source UEs, instead of sharing radio signals, as shown in Fig. 3. Using this small amount of shared information, channel estimation processing functions (CEFs) for inter-site interference are deployed for each site. CEF demodulates the pilot signal of the interfering source UE using the pilot assignment information and estimates channels for inter-site interference independently for each site. By suppressing the inter-site interference using this estimated channel information from CEF independently for each site, the proposed method provides the same inter-site interference reduction effect as that obtainable by transmitting radio signals and maintains the wireless communication quality with a lower transmission load between sites.

In STEP 1 of Fig. 3, which shows the operation flow, the uRIC assigns the pilot allocations and calculates AP cluster $\boldsymbol{D}_k$ and the interference AP list for each UE using the power information measured at each AP. Here, the interference AP list is the set of APs whose signal power from the UE exceeds the threshold and includes AP cluster $\boldsymbol{D}_k$. Let interference cluster $\boldsymbol{E}_k$ belonging to an interference AP list be defined as the following L-dimensional square matrix

$$\boldsymbol{E}_k = \begin{bmatrix} E_{k1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & E_{kL} \end{bmatrix}, \tag{12}$$

where, E_{kl} is defined as

$$E_{kl} = \begin{cases} 1 & \text{if UE } k \text{ causes interference to AP } l, \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$



STEP 1. uRIC assigns the pilot allocations and calculates AP clusters and interference clusters per UE, and shares the calculation results with the respective sites.

STEP 2. CEF at each site performs channel estimation and shares the estimated channel with vCPU.

STEP 3. vCPU at each site calculates weight and performs receive signal processing.

FIGURE 3. Flow of the proposed CPU cooperation method.

The pilot allocations, $\boldsymbol{D}_k$ and $\boldsymbol{E}_k$, for each user are updated in accordance with user mobility and shared with the site where the CPUs connected by the user are deployed. In STEP 2 of Fig. 3, vCPUs for signal processing for each UE and CEFs for channel estimation of inter-site interference signals are deployed at each site. The CEF in site j performs channel estimation for APs connected to site j among APs in the interference cluster of UEs whose vCPU is deployed other than in site j. The CEF is a network function that serves as a channel estimator of inter-site interference. In STEP 3 of Fig. 3, the vCPU calculates the weights in (4) using the estimated channel information of inter-site interference from the CEF. This coherent signal processing can suppress inter-site interference.

Note that interference cluster $\boldsymbol{E}_k$ is not defined in P-MMSE in [8], as $\boldsymbol{E}_k$ is treated equivalently to AP cluster $\boldsymbol{D}_k$. The interference cluster $\boldsymbol{E}_k$ is the set of APs that perform channel estimation for UE k, which differs from the AP clusters. In the proposed method, the channel estimation process by the CEF generates a pair of UE and AP in which the UE's radio signals are not processed for transmission and reception, but channel estimation using only pilot signals is performed instead. Therefore, interference cluster $\boldsymbol{E}_k$ is defined as a new AP set different from the AP cluster. In the mathematical expression, it is equivalent to extending $\mathcal{P}_k$ in (4) as $\mathcal{P}_k = \{i : \boldsymbol{D}_k\boldsymbol{E}_i \neq \boldsymbol{O}_L\}$.

In the example of Fig.3, the radio signals of the AP clusters of UE#1 and UE#2 are aggregated at edge site#1 and edge site#2, respectively. The signal processing of UE#1 and UE#2 is performed in vCPU#1 and vCPU#2, respectively. CEF#1 performs channel estimation between UE#2 and AP#3,4 included in the interference cluster of UE#2, whose vCPU is not deployed in Site#1. The channel for the inter-site interference estimated by the CEF is shared with the vCPUs in the same site. The inter-site interference can be suppressed by calculating the weights in (4) that substitute the estimated channel for the inter-site interference. The CEF performs signal processing for channel estimation independently at each site, so the estimated channels and radio signals of the interference signals between APs and UEs estimated by the CEF are not shared between the sites. The uRIC shares the AP cluster $\boldsymbol{D}_k$, pilot assignment information, and interference cluster $\boldsymbol{E}_k$ with each site. In the proposed interference suppression method, the amount of data to be shared is $2KL+K\tau$ bits, where τ is the pilot sequence length. This data volume is significantly smaller than that of the radio signals, making the proposed method more efficient in terms of transmission load compared to the existing method that aggregates radio signals. Therefore, the proposed method can suppress inter-site interference and achieve high radio quality with a lower transmission load.

B. RAN MANAGEMENT FOR PROPOSED METHOD

Next, we consider the formation of the interference cluster in user-centric RAN using the proposed method. This formation of the interference cluster generates the minimum transmission load and imposes a lower computational load for the required radio quality in the area. The increase in the number of APs in the interference cluster, which is a new metric in the proposed method, can suppress inter-site interference and not affect the transmission load on the BH but increases the computational load according to (10), (11). Therefore, the size of interference clusters needs to be managed to balance the computational load and interference suppression effect. In addition, the size of the AP cluster affects radio quality, transmission load, and computation load according to (6), (10), and (11). Therefore, the size of the interference cluster along with the AP cluster needs to be managed by the uRIC

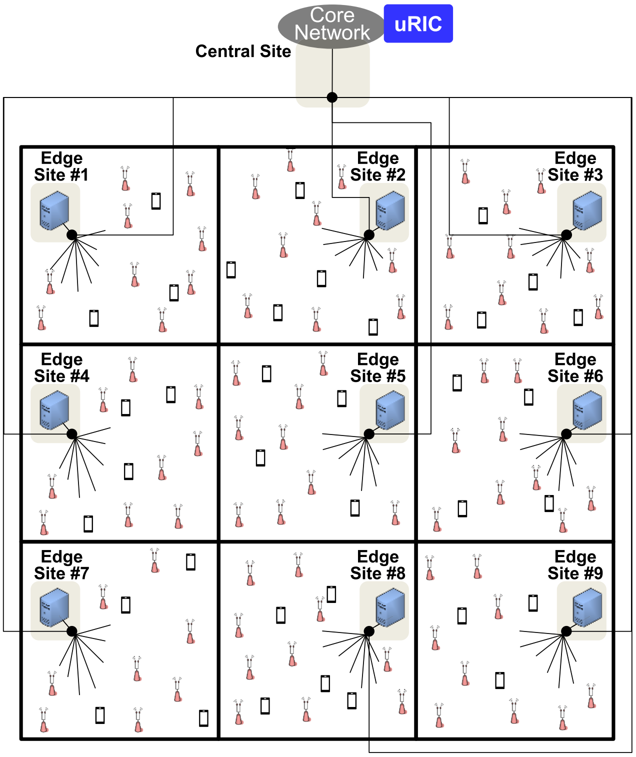


FIGURE 4. Evaluation environment. The figure shows an example where the number of edge sites is four ($J = 9$).

TABLE 1. Simulation parameters.

Parameter	Value
Target area size	1 km × 1 km
The number of edge sites, J	9
The number of APs, L	400
Placement of APs and UEs	Random with uniform distribution
The number of antenna per AP	1
The number of UEs, K	100
User traffic load	Full buffer
Carrier frequency, Bandwidth	3.5 GHz, 100 MHz
Transmission power of UE	100 mW
Subcarrier spacing	30 kHz
Number of pilot sequences	36
Quantization bit rate	16 bits/symbol
Path loss	3GPP-UMi [34]
Channel fading	Correlated Rayleigh fading [35]
Channel estimation	MMSE [17]
Shadowing deviation	10 dB
Noise figure	7 dB
KPI_C, KPI_{TP}	15 GFlops, 300 Mbps
GA parameters -Population size -Number of generations -Mutation rate	 - 50 - 200 - 0.2

based on the user centric-RAN architecture, which assumes the actual physical structure for balancing the radio quality, transmission line load, and computational load.

To manage the size of AP cluster $\boldsymbol{D}_k$ and interference cluster $\boldsymbol{E}_k$, we adopt the following AP cluster formation algorithm proposed in [8].

- When UE k initiates communication, it sends a request to the AP with the best channel status. The AP that receives the request becomes the master AP for UE k. The master AP services UE k and assigns the pilot with the least pilot contamination to UE k at that time.
- For all APs that can be selected for the AP cluster for UE k, APs are associated with the cluster if the difference in channel gain between the master AP of UE k and the AP is within x dB.

Interference clusters are calculated in the same way using the following algorithm. For all APs selectable for the interference cluster for UE k, APs are associated with the cluster if the difference in channel gain between the master AP of UE k is within y dB. This rule-based algorithm has low complexity and provides uniform radio quality. However, in [8], the threshold value x related to the size of AP clusters is fixed as a parameter, and management according to the computer and transmission resources in RAN has not been implemented. In this paper, the threshold x for AP clusters and the threshold y for interference clusters are treated as decision variables to be calculated for each area. Here, the optimization problem of balancing radio quality and computer and transmission load in user-centric RAN can be described as follows.

$$\min_{x,y} \; R_{\mathrm{BH}}, \tag{14a}$$

$$\text{s.t.}: \frac{1}{K}\sum_{k\in K} TP_k \geq KPI_{\mathrm{TP}}, \tag{14b}$$

$$C_{comp} \leq KPI_C, \tag{14c}$$

$$x \leq y. \tag{14d}$$

The objective function (14a) minimizes the transmission load on the BH. The reduction in transmission load allows the MNO to install the minimum number of switches and transponders for the required wavelength division multiplexing, which leads to cost reduction. In constraint (14b), KPI_{TP} is the index to be satisfied for the average user throughput of the area and is set as the lower limit of the radio quality provided by the MNO. In constraint (14c), KPI_C is set by the MNO as an upper bound on the amount of processing to be allocated to radio signal processing for the commodity servers deployed at each site.

By solving the optimization problem (14), we can determine the threshold values x and y of the AP and interference clusters that minimize the transmission line load of the BH while satisfying the radio quality KPI_{TP} and the computer load KPI_C. In other words, we can achieve the required radio quality with the minimum transmission load by managing the size of AP clusters and interference clusters for each area. Thus, we can solve the trade-off problem between transmission line load and radio quality. As the optimization problem (14) has a non-linear objective function and constraints, it is solved using a genetic algorithm (GA) [31]. A GA is a metaheuristic inspired by the process of natural selection that belongs to the larger class of evolutionary algorithms. It is used to find global minima for non-linear optimization problems at various layers of a communication system. According to [32], the complexity of the genetic algorithm is on the order of $O(gpq)$ where g is the number of generations, p is the population size, and q is the size of the individuals. Here, q is of the same order as $\sum_{k \in K} C_{\text{decode}}(k)$. Considering user mobility, it is necessary to reduce the computational complexity of the search for a quasi-optimal solution because the number of iterations required to solve the optimization problem is increasing. A partially modified GA method [33] has also been proposed to reduce the computational complexity, and the application of such method will be the subject of our future work.

IV. PERFORMANCE EVALUATION

In this section, we present the numerical results of a computer simulation to demonstrate the performance of the proposed method.

A. EVALUATION CONDITIONS

This section describes the conditions in the evaluation. Fig. 4 shows the evaluation environment in which CPUs are distributed in each area, assuming the deployment of a user-centric RAN with CF-mMIMO in urban areas. The user-centric RAN architecture is based on one central site that covers the area, and $J = 9$ edge sites are placed in a grid pattern to divide the area into 9 sections. Each AP is connected to an optical line terminal at the edge site of the deployed divided area, and the edge site is connected to the optical line terminal at the central site via optical fiber. Table 1 presents the simulation conditions used in the computer simulation. In an area of 1 km^2, 400 APs and UEs are randomly placed. The channel quality is calculated using the 3GPP UMi model [34] for path loss and shadowing, and correlated Rayleigh fading [35]. Since the terminal is stationary, the calculation is repeated ten times to enable statistical evaluation. Number of pilot sequences is 36, so there are UEs whose assigned pilots are not orthogonal to each other.

B. EVALUATION RESULTS AND ANALYSIS

To evaluate the performance of the proposed method, we compare the proposed method with the existing method, which involves aggregating radio signals at the CPU as presented in [10]; we refer to this as "Existing CPU cooperation". In addition, we compare the proposed method with a method that applies coordinated beamforming (CoBF) techniques [29], [30] to CF-mMIMO. This suppresses inter-cell interference between cells at different sites. Here, "CF-mMIMO with CoBF" indicates the results when a coordinated beamforming (CoBF) technique is simply applied to CF-mMIMO without the proposed management of the formation of interference clusters. For the existing CPU cooperation, the method based on [8] is adopted to allow a fair comparison.

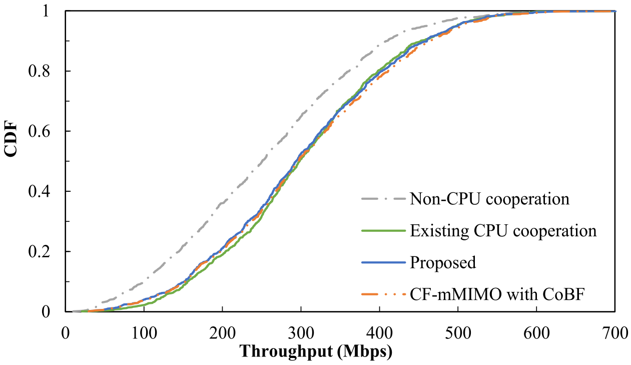


FIGURE 5. Comparison of the CDF of user throughput.

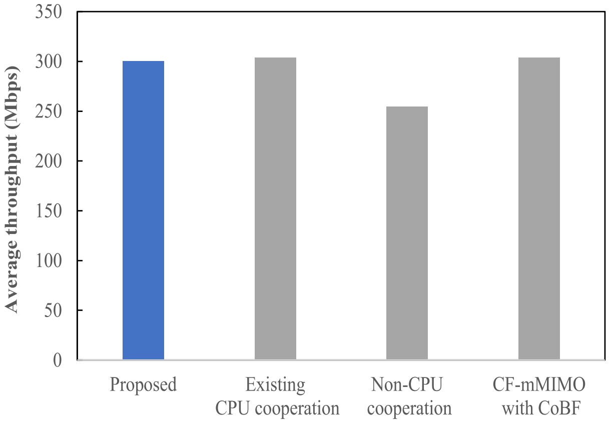


FIGURE 6. Comparison of the average user throughput.

In the evaluation, the following three factors are compared based on the problems described in Section II. We compare the uplink throughput to evaluate the radio quality in terms of user experience. Next, we compare the computational load and the uplink transmission load from the perspective of RAN resource utilization.

First, we show the simulation results of the cumulative distribution function (CDF) and the average value of user throughput in Fig. 5 and Fig. 6, respectively, to confirm the effect of suppression of inter-site interference. Here, "Non-CPU cooperation" is a method with no inter-CPU cooperation and no suppression of inter-site interference as presented in [6]. AP clusters are also determined by the algorithm presented in [8], but AP clusters are not formed across sites. The results show that the proposed method, the existing cooperation method, and CF-mMIMO with CoBF are almost equivalent, slightly exceeding the average user throughput. This is because these methods form the minimum size AP cluster that achieves the throughput KPI by optimizing the AP cluster and/or interference cluster as the decision variables. On the other hand, we can also see that the throughput of the non-CPU cooperation method is 15% lower on average and, in particular, 42% lower for the 5%-tile value than that of the method with cooperation. This is because the non-cooperation method cannot suppress the inter-site interference, which degrades the radio quality compared to the method with cooperation.

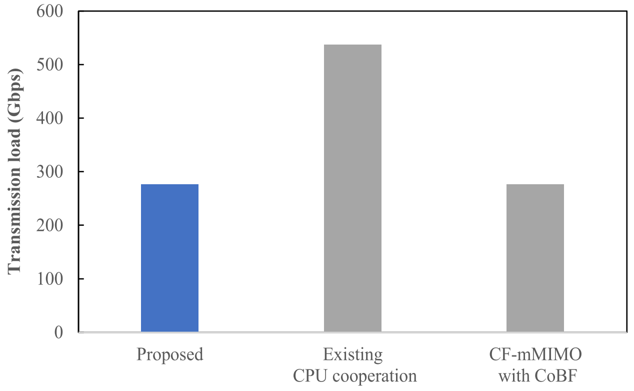


FIGURE 7. Comparison of the average transmission load in BH.

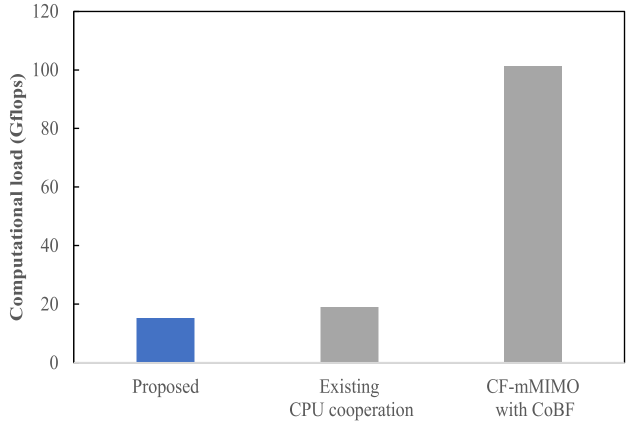


FIGURE 8. Comparison of the average computational load.

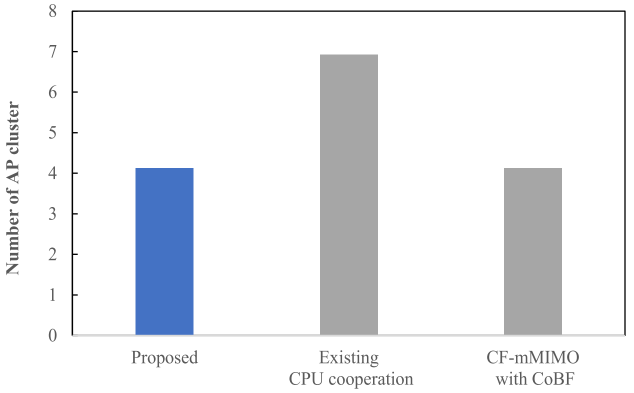


FIGURE 9. Comparison of the average number of APs forming AP clusters.

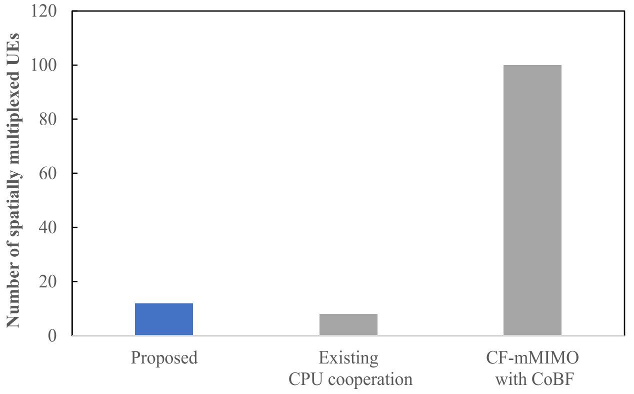


FIGURE 10. Comparison of the average number of UEs to be spatially multiplexed.

Figs. 7 and 8 present the transmission and computational loads, respectively. As shown in Fig. 7, the transmission load of the proposed method and CF-mMIMO with CoBF is reduced by 53% compared to the existing cooperation method. Fig. 8 shows that the proposed method reduces the computational load by 15% compared to the existing cooperation method and by 83% compared to CF-mMIMO with CoBF. Therefore, we can confirm that the proposed method provides high radio quality by suppressing inter-site interference while significantly reducing the transmission load, particularly in BH, as well as the computational load.

We next explain the reasons for the reduction in transmission load achieved by the proposed method. Fig. 9 shows a comparison of the average number of APs in the AP cluster. From Fig. 9, we can see that the number of APs forming an AP cluster is higher in the existing cooperation method than in the proposed method. The existing cooperation method suppresses inter-site interference by expanding the AP clusters across the sites, but the transmission load is high due to the aggregation of radio signals via BH. By forming AP clusters with a smaller number of APs, the proposed method can reduce the number of inter-site APs forming AP clusters and decrease the amount of radio signals transmitted via BH. To analyze why the proposed method provides high radio quality even with small AP clusters, Fig. 10 shows $\mathcal{P}_k$ indicating the spatially multiplexed UEs in the weight calculation. From Fig. 10 we can see that the number of spatially multiplexed UEs $\mathcal{P}_k$ is larger in the proposed method. This indicates that the proposed method can perform independent channel estimation of inter-site interference for each site by CEF in a wider range without increasing the transmission line load and suppress inter-site interference using P-MMSE in (4). Therefore, the proposed method can achieve the same radio quality as the existing cooperative method with fewer APs in AP clusters by increasing the number of UEs to be spatially multiplexed by expanding the interference cluster through solving the optimization problem (14). Next, we explain the reasons for the reduction in computational load achieved by the proposed method. From Fig. 9, CF-mMIMO with CoBF increases the size of interference clusters due to the lack of proper management of these clusters, resulting in increased computational load. On the other hand, as shown in Fig. 9 and Fig. 10, the proposed method can balance the reduction in the computational load caused by the shrinking of the AP clusters and the increase in computational load caused by the expansion of the interference cluster using the optimization (14).

To verify the proposed method's versatility, the transmission load in BH for different numbers of UEs is presented in Fig. 11. It can be observed that the proposed method achieves

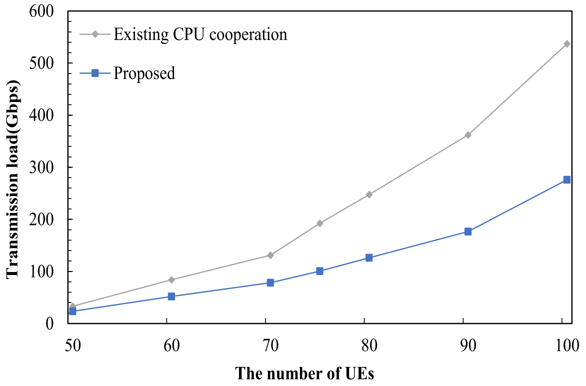


FIGURE 11. Comparison of the average transmission load for the number of UEs.

a lower transmission load than the existing cooperation method, regardless of the number of UEs. This result demonstrates the effectiveness of the proposed method in reducing the transmission load while maintaining high radio quality in various scenarios. Also, we can see that the transmission load increases as the number of UEs increases regardless of the method. This is because the size of the AP cluster expands to suppress the interference caused by the increase in the number of UEs to achieve the user throughput KPI. This expansion of AP clusters increases the number of AP clusters across the sites, which also causes the transmission load to increase. It can be seen from Fig. 11 that the gap in transmission load for each method becomes increasingly large as the number of UEs increases. As the number of UEs increases, the size of AP clusters expands to achieve KPI_{TP}. In the existing cooperation method, the expansion in AP cluster size causes more APs from other sites to include the AP cluster, thus increasing the transmission load between sites. On the other hand, the proposed method can increase throughput while reducing the spread of AP clusters by expanding the interference clusters.

V. CONCLUSION

We proposed an inter-site CPU cooperation method that maintains high radio quality while reducing the transmission load on the backhaul connecting the sites, which is necessary to suppress inter-site interference. The proposed method achieves inter-site interference suppression by sharing a list of APs causing interference and pilot assignment information among sites, instead of sharing radio signals as in the existing CPU cooperation method. This approach significantly reduces the transmission load. The proposed method also performs independent channel estimation and interference suppression processes site-by-site, which reduces the same inter-site interference as that of shared radio signals. Furthermore, we introduced optimization management that adjusts the level of CPU cooperation based on the proposed method to minimize transmission load and computational load while providing the required radio quality in the area.

The simulation results demonstrate that the proposed method significantly reduces the transmission load on BHs by approximately 53% compared to the existing scheme when the radio quality and computational load in the area are similar. This is because the proposed method estimates more inter-site interference and increases the number of UEs that can be multiplexed by P-MMSE. This approach enables the proposed method to achieve the same radio quality as the existing cooperative scheme even if the number of APs in the AP cluster is small.

In a future work, we plan to extend and evaluate our method to achieve both high communication quality and low transmission load, considering user mobility and a mixed environment of UEs with different required communication qualities.

ACKNOWLEDGMENT

An earlier version of this paper was presented at the conference papers [36], [37].

REFERENCES

[1] 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.

[2] KDDI. (May 2021). *Beyond 5G/6G White Paper*. [Online]. Available: https://www.kddi-research.jp/english/tech/whitepaper_b5g_6g/assets/pdf/KDDI_B5G6G_WhitePaperEN_2.0.1.pdf

[3] H. Q. Ngo, A. Ashikhmin, H. Yang, E. G. Larsson, and T. L. Marzetta, "Cell-free massive MIMO: Uniformly great service for everyone," in *Proc. IEEE 16th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Stockholm, Sweden, Jun. 2015, pp. 201–205.

[4] X. Li, X. Zhang, Y. Zhou, and L. Hanzo, "Optimal massive-MIMO-aided clustered base-station coordination," *IEEE Trans. Veh. Technol.*, vol. 70, no. 3, pp. 2699–2712, Mar. 2021.

[5] M. Matthaiou, O. Yurduseven, H. Q. Ngo, D. Morales-Jimenez, S. L. Cotton, and V. F. Fusco, "The road to 6G: Ten physical layer challenges for communications engineers," *IEEE Commun. Mag.*, vol. 59, no. 1, pp. 64–69, Jan. 2021.

[6] C. D'Andrea, G. Interdonato, and S. Buzzi, "User-centric handover in mmWave cell-free massive MIMO with user mobility," in *Proc. 29th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2021, pp. 1–5.

[7] G. Interdonato, P. Frenger, and E. G. Larsson, "Scalability aspects of cell-free massive MIMO," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–6.

[8] E. Björnson and L. Sanguinetti, "Scalable cell-free massive MIMO systems," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4247–4261, Jul. 2020.

[9] V. Ranjbar, A. Girycki, M. A. Rahman, S. Pollin, M. Moonen, and E. Vinogradov, "Cell-free mMIMO support in the O-RAN architecture: A PHY layer perspective for 5G and beyond networks," *IEEE Commun. Standards Mag.*, vol. 6, no. 1, pp. 28–34, Mar. 2022.

[10] C. D'Andrea and E. G. Larsson, "User association in scalable cell-free massive MIMO systems," in *Proc. 54th Asilomar Conf. Signals, Syst., Comput.*, Nov. 2020, pp. 826–830.

[11] F. Riera-Palou and G. Femenias, "Decentralization issues in cell-free massive MIMO networks with zero-forcing precoding," in *Proc. 57th Annu. Allerton Conf. Commun., Control, Comput. (Allerton)*, Sep. 2019, pp. 521–527.

[12] H. A. Ammar, R. Adve, S. Shahbazpanahi, G. Boudreau, and K. V. Srinivas, "Distributed resource allocation optimization for user-centric cell-free MIMO networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 5, pp. 3099–3115, May 2022.

[13] K. Yamazaki, T. Ohseki, Y. Amano, H. Shinbo, T. Murakami, and Y. Kishi, "Proposal for a user-centric RAN architecture towards beyond 5G," in *Proc. ITU Kaleidoscope, Connecting Phys. Virtual Worlds (ITU K)*, Dec. 2021, pp. 1–7.

[14] L. Bonati, M. Polese, S. D'Oro, S. Basagni, and T. Melodia, "Open, programmable, and virtualized 5G networks: State-of-the-art and the road ahead," *Comput. Netw.*, vol. 182, Dec. 2020, Art. no. 107516.

[15] M. Peng, Y. Sun, X. Li, Z. Mao, and C. Wang, "Recent advances in cloud radio access networks: System architectures, key techniques, and open issues," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 2282–2308, 3rd Quart., 2016.

[16] P. Chanclou, H. Suzuki, J. Wang, Y. Ma, M. R. Boldi, K. Tanaka, S. Hong, C. Rodrigues, L. A. Neto, and J. Ming, "How does passive optical network tackle radio access network evolution?" *J. Opt. Commun. Netw.*, vol. 9, no. 11, pp. 1030–1040, Nov. 2017.

[17] T. H. Nguyen, T. K. Nguyen, H. D. Han, and V. D. Nguyen, "Optimal power control and load balancing for uplink cell-free multi-user massive MIMO," *IEEE Access*, vol. 6, pp. 14462–14473, 2018.

[18] E. Nayebi, A. Ashikhmin, T. L. Marzetta, and B. D. Rao, "Performance of cell-free massive MIMO systems with MMSE and LSFD receivers," in *Proc. 50th Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, USA, Nov. 2016, pp. 203–207.

[19] E. Björnson and L. Sanguinetti, "Making cell-free massive MIMO competitive with MMSE processing and centralized implementation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 77–90, Jan. 2020.

[20] *Small Cell Forum, Small Cell Virtualization: Functional Splits and User Cases*, Jan. 2016.

[21] Y. Tsukamoto, R. K. Saha, S. Nanba, and K. Nishimura, "Experimental evaluation of RAN slicing architecture with flexibly located functional components of base station according to diverse 5G services," *IEEE Access*, vol. 7, pp. 76470–76479, 2019.

[22] H. A. Ammar, R. Adve, S. Shahbazpanahi, G. Boudreau, and K. V. Srinivas, "User-centric cell-free massive MIMO networks: A survey of opportunities, challenges and solutions," *IEEE Commun. Surveys Tuts.*, vol. 24, no. 1, pp. 611–652, 1st Quart., 2022.

[23] Z. H. Shaik, E. Björnson, and E. G. Larsson, "MMSE-optimal sequential processing for cell-free massive MIMO with radio stripes," *IEEE Trans. Commun.*, vol. 69, no. 11, pp. 7775–7789, Nov. 2021.

[24] Ö. T. Demir and E. Björnson, "Max-min fair wireless-powered cell-free massive MIMO for uncorrelated Rician fading channels," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, May 2020, pp. 1–6.

[25] C. Wei, K. Xu, X. Xia, Q. Su, M. Shen, W. Xie, and C. Li, "User-centric access point selection in cell-free massive MIMO systems: A game-theoretic approach," *IEEE Commun. Lett.*, vol. 26, no. 9, pp. 2225–2229, Sep. 2022.

[26] V. Ranasinghe, N. Rajatheva, and M. Latva-aho, "Graph neural network based access point selection for cell-free massive MIMO systems," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Madrid, Spain, Dec. 2021, pp. 1–6.

[27] Q. N. Le, V.-D. Nguyen, O. A. Dobre, N.-P. Nguyen, R. Zhao, and S. Chatzinotas, "Learning-assisted user clustering in cell-free massive MIMO-NOMA networks," *IEEE Trans. Veh. Technol.*, vol. 70, no. 12, pp. 12872–12887, Dec. 2021.

[28] *O-RAN Architecture Description*, O-RAN Alliance, Buschkaulerweg, Germany, Nov. 2020.

[29] M. K. Karakayali, G. J. Foschini, and R. A. Valenzuela, "Network coordination for spectrally efficient communications in cellular systems," *IEEE Wireless Commun.*, vol. 13, no. 4, pp. 56–61, Aug. 2006.

[30] S. Lakshminarayana, M. Assaad, and M. Debbah, "Coordinated multicell beamforming for massive MIMO: A random matrix approach," *IEEE Trans. Inf. Theory*, vol. 61, no. 6, pp. 3387–3412, Jun. 2015.

[31] K. Goudos, "Evolutionary algorithms for wireless communications—A review of the state-of-the art," in *Contemporary Issues in Wireless Communications*. Rijeka, Croatia: InTech, 2014.

[32] B. Rylander and J. A. Foster, "Computational complexity and genetic algorithms," Tech. Rep., 2001.

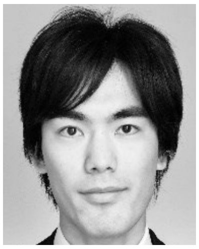
[33] B. Makki, A. Ide, T. Svensson, T. Eriksson, and M.-S. Alouini, "A genetic algorithm-based antenna selection approach for large-but-finite MIMO networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 7, pp. 6591–6595, Jul. 2017.

[34] *Study on Channel Model for Frequencies From 0.5 to 100 GHz*, 3GPP, document TR 38.901 V16.1.0, Dec. 2019.

[35] E. Björnson, J. Hoydis, and L. Sanguinetti, "Massive MIMO networks: Spectral, energy, and hardware efficiency," *Found. Trends® Signal Process.*, vol. 11, nos. 3–4, pp. 154–655, 2017.

[36] A. Ikami, Y. Tsukamoto, N. Aihara, T. Murakami, and H. Shinbo, "Interference suppression for distributed CPU deployments in cell-free massive MIMO," in *Proc. IEEE 96th Veh. Technol. Conf.*, Sep. 2022, pp. 1–6.

[37] T. Murakami, N. Aihara, A. Ikami, Y. Tsukamoto, and H. Shinbo, "Analysis of CPU placement of cell-free massive MIMO for user-centric RAN," in *Proc. IEEE/IFIP Netw. Oper. Manage. Symp.*, Budapest, Hungary, Apr. 2022, pp. 1–7.



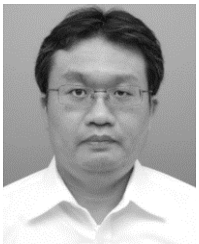
AKIO IKAMI received the B.E. degree in electrical and electronic engineering from Kyoto University, Kyoto, Japan, in 2011, and the M.Sc. degree in communications and computer engineering from the Graduate School of Informatics, Kyoto University, in 2013. He joined KDDI Corporation, Tokyo, Japan, in 2013, and became engaged in the development of access networks. Since 2018, he has been with KDDI Research, Inc., Saitama, Japan. His research interests include RAN management and optimization for 6G. He received the Young Researcher's Award from IEICE, in 2020, and the Best Paper Award from the IEEE VTC2022 Fall.

NAOKI AIHARA received the B.E. and M.E. degrees from The University of Electro-Communications, Tokyo, Japan, in 2018 and 2020, respectively.

He joined KDDI Corporation, in 2020. Since 2021, he has been a member of KDDI Research, Inc., Saitama, Japan. His research interests include wireless communication systems, such as radio access networks and application of machine learning to wireless communication. He was a recipient of the IEICE RCS 2019 Active Research Award. He was a co-recipient of the WPMC 2020 Best Student Paper Award.

YU TSUKAMOTO received the A.E. degree in electronic control system engineering from the National Institute of Technology, Numazu College, Shizuoka, Japan, in 2012, and the B.E. degree in electrical engineering from the Kyoto Institute of Technology, Kyoto, Japan, in 2014, and the M.E. degree in electrical engineering from Kyoto University, Kyoto, in 2016. He joined KDDI Corporation, Tokyo, Japan, in 2016. Since 2017, he has been a member of KDDI Research, Inc., Saitama, Japan, where he studies radio access network technologies.

TAKAHIDE MURAKAMI received the B.E., M.E., and Ph.D. degrees in communication engineering from Tohoku University, Sendai, Japan, in 2002, 2004, and 2007, respectively. He joined KDDI Corporation, in 2007. He is currently a member of KDDI Research, Inc., (formerly KDDI Research and Development Laboratories). His research interests include wireless communications and radio access networks.



HIROYUKI SHINBO received the B.S. degree in electro information communication from The University of Electro-Communications, Tokyo, Japan, in 1987, and the M.S. degree from the Graduate School of Information and System, The University of Electro-Communications, in 1990. He joined KDD Corporation (now KDDI Corporation), in 1999. He was with KDD Research and Development Laboratories (now KDDI Research, Inc.). He was seconded to work with the Advanced Telecommunications Research Institute International, from 2013 to 2016. He is currently a Senior Manager of the Advanced Radio Application Laboratory, KDDI Research, Inc. His research interests include beyond 5G/6G systems (especially radio access networks), TCP/IP, flying base stations, network operation systems, and space communications.