Fronthaul Load-Reduced Scalable Cell-Free massive MIMO by Uplink Hybrid Signal Processing

Issei Kanno*, Masaaki Ito*†, Takeo Ohseki*, Kosuke Yamazaki*, Yoji Kishi*,
Thomas Choi†, Wei-Yu Chen†, and Andreas F. Molisch†
* KDDI Research Inc., Saitama, Japan
†University of Southern California, Los Angeles, United States

***Abstract*—This paper proposes hybrid signal processing schemes for the uplink cell-free massive MIMO; these schemes serve to reduce fronthaul loads to obtain a scalable centralized processing architecture. In this architecture, received signals of multiple receive antennas at the access points (APs) are compressed into fewer streams by local spatial signal processing and then the streams are forwarded to a central processing unit (CPU) via fronthaul, and the CPU performs scalable processing for channel estimation and signal detection based on partial minimum mean squared error (PMMSE). We propose two kinds of concrete local signal processing methods for this hybrid processing architecture: one is based on MMSE, and the other is based on principal component analysis (PCA) with eigenvalue decomposition (EVD). For the EVD, a local vector selection based EVD (LVS-EVD) that selects uniform number of eigenvectors for each AP in a standalone way, and a global vector selection based EVD (GVS-EVD) that determines the dimensions of the weight vector of each AP in the CPU, are further considered. Computer simulations verify the approaches and compare their effectiveness. In addition, we show that the GVS-EVD scheme can be operated with significantly reduced fronthaul loads without severe performance degradation.**

***Index Terms*—Scalable cell-free massive MIMO, Uplink hybrid signal processing, Fronthaul loads reduction, MMSE, EVD**

I. Introduction

Cell-free massive MIMO (CFmMIMO) is a promising technology to enhance capacity, reliability and energy efficiency [1], [2], and can be considered as a candidate architecture for future mobile communication systems such as beyond 5G [3]. One critical issue to deploy this technology over wide service area is to have a scalability [4], meaning limited computational complexity for the signal processing, access control and so on. In [5], two types of scalable signal processing schemes have been proposed, and shown that they can realize uplink signal detection, downlink precoding and channel estimation for each user efficiently without increasing the amount of computational complexity per users significantly by employing dynamic cooperation clustering for the signal processing. One is central processing where the received signals of all APs are hauled to the CPU, and CPU operates partial MMSE (PMMSE). The other is a semi-distributed architecture, which applies local PMMSE (LP-MMSE) at the AP side and combines the result in the CPU. Although PMMSE requires higher computational complexity than LPMMSE, it can achieve higher spectral efficiency. However, a key remaining challenge is that the fronthaul capacity increases with the number of APs and the number of antennas of each AP. Given the fact that fronthaul loads are anticipated to increase in beyond 5G systems compared to the current values because of increases in the bandwidth of the wireless access links, efficient schemes to reduce the fronthaul load will be desirable.

In this paper, we focus on the reduction of the fronthaul load of scalable central processing for uplink CFmMIMO. Ref. [6] analyzes the effect of the limited backhaul capacity for uplink and discusses suitable transmission power control and combining filter design for maximum ratio combining (MRC) receiver, and [7] combines AP selection to compute and forward structure. However, while MRC type receivers allow simple computations, their spectral efficiency is inferior to MMSE type receivers [8], [9]. Hence, in this paper we propose a hybrid processing architecture suitable for the PMMSE in order to reduce the fronthaul load. In this architecture, the received signals of multiple receive antennas at the APs are compressed into fewer streams by a spatial signal processing and then the streams are forwarded to the CPU via fronthaul. The CPU performs channel estimation and signal detection by PMMSE. By designing the number of output ports smaller than the number of antenna ports of each AP, this can effectively reduce the fronthaul load. Concretely, two specific spatial processing schemes are applied to the local processing of the AP. One is based on MMSE, for which weight vectors of the local processing are designed to reduce interference from other UEs to extract signals from specific UEs that have smaller pathloss. The other is based on PCA with EVD of the covariance matrix of the received signal vector of each AP. This sets the eigenvectors with larger eigenvalue as weight vectors of the local processing to extract the principal components at each AP. For selecting the eigenvectors, LVS-EVD that selects a uniform number of eigenvectors in each AP and a GVS-EVD that determines the dimensions of the weight vector of each AP adaptively in the CPU are considered. Computer simulations in Sec. IV show that the hybrid processing with these schemes can reduce the fronthaul loads effectively. In addition, it shows the hybrid processing with GVS-EVD can be operated with significantly lower fronthaul loads without severe degradation from conventional PMMSE.

II. System Model

Consider an uplink CFmMIMO system with L APs with N antennas each, and K single-antenna UEs that are spatially

multiplexed. The uplink received signal vector of the l-th AP for the i-th symbol is expressed as follows,

$$\mathbf{r}_l(i) = \mathbf{H}_l \mathbf{s}(i) + \mathbf{n}_l(i), \tag{1}$$

where $\mathbf{H}_l$ is an N x K channel matrix of the l-th AP, of which the (n,k)-th element, $h_{n,k}^{(l)}$ is the channel response between the k-th UE and the n-th antenna, and can be expressed as

$$h_{n,k}^{(l)} = \sqrt{\beta_{l,k}} p_{n,k}^{(l)}, \tag{2}$$

where $\beta_{l,k}$ describes the path gain and large-scale fading between APs and UEs, and $p_{n,k}^{(l)}$ the small-scale fading, which additionally depends on the considered antenna element at that AP. We assume here frequency-flat channels though generalization to OFDM can be done. $\mathbf{s}(i)$ is a K dimensional transmitted symbol vector of which the k-th element is the signal of the k-th UE $s_k(i)$, and it assumes all the UEs transmit signals with the equal transmission power. $\mathbf{n}(i)$ is a N dimensional noise vector, of which n-th element is that of the n-th receive antenna. In the centralized architecture of CFmMIMO, all the received signals of APs are forwarded to the CPU via fronthaul, and the purpose of the signal detection is to detect $s_k(i)$ for all UEs from the received signals. The CPU can also be utilized for channel estimation of each UE and for generating the combining weight for the detection. The received signal vectors at the CPU can be written as $\mathbf{y}(i) = \mathbf{H}\mathbf{s}(i) + \mathbf{n}(i)$, where $\mathbf{r}(i) = [\mathbf{r}_1(i), \mathbf{r}_2(i), \cdots \mathbf{r}_L(i)]^{\mathrm{T}}$, $\mathbf{H} = [\mathbf{H}_1^{\mathrm{T}}, \mathbf{H}_2^{\mathrm{T}}, \cdots \mathbf{H}_L^{\mathrm{T}}]^{\mathrm{T}}$and $\mathbf{n}(i) = [\mathbf{n}_1(i), \mathbf{n}_2(i), \cdots \mathbf{n}_L(i)]^{\mathrm{T}}$, respectively. Note that it is assumed that the fronthaul forwards received signal at sufficient quantization level and the effect of the quantization noise is negligible.

In PMMSE, the weight vector $\mathbf{v}_k$ to detect the signal of the UE k can be written as follows,

$$\mathbf{v}_k = (\sum_{i \in \mathcal{D}_k} \mathbf{D}_k \hat{\mathbf{h}}_i \hat{\mathbf{h}}_i^{\mathrm{H}} \mathbf{D}_k + \frac{1}{\rho} \mathbf{D}_k)^{\dagger} \mathbf{D}_k \hat{\mathbf{h}}_k, \tag{3}$$

where $\mathbf{h}_k$ is the k-th column vector of the matrix $\mathbf{H}$, and $\hat{\mathbf{h}}_k$ is its estimate. $\mathbf{D}_k$ is a clustering matrix to extract the corresponding elements of UE k, and it can be written as $\mathbf{D}_k = diag[\mathbf{D}_{k1}, \mathbf{D}_{k2}, \cdots \mathbf{D}_{kL}]$. $\mathbf{D}_{kl}$ expresses the relationship between the UE k and AP l, as follows; if the AP l is a member of the cluster for UE k, $\mathbf{D}_{kl} = \mathbf{I}_N$, and otherwise $\mathbf{0}_N$. † denotes the generalized inverse operation. The set $\mathcal{D}_k$ is defined as $\mathcal{D}_k = \{i, \mathbf{D}_i \mathbf{D}_k \neq \mathbf{0}\}$. ρ is the transmission power to noise ratio. Then the detected signal of UE k can be expressed as $\hat{s}_k(i) = \mathbf{v}_k^{\mathrm{H}} \mathbf{r}(i)$. In [5], it is discussed that by utilizing a clustering scheme, the actual computational complexity of the CPU to detect each UE does not grow with the number of APs and UEs. For selection of the APs for each UE k, various clustering schemes have been proposed [10], [11]. In this paper, we simply assume each UE selects a uniform number of APs, defined as Z, with higher large-scale fading factor $\beta_{l,k}$. For example, it can be measured if each AP periodically transmits synchronization signals that are orthogonal with each other, and associated with cell ID of each cell in a 3GPP system. The CPU can determine the APs for each UE if it assumes the UEs report the measured large-scale fading data periodically.

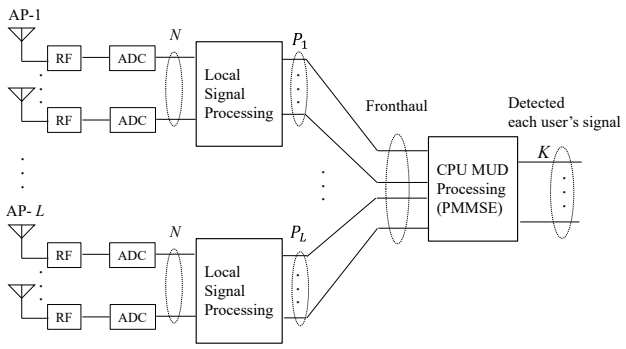


Fig. 1. Architecture of proposed receiver.

The fronthaul is required to forward NQ bits from each AP to the CPU and LNQ bits totally in this centralized architecture, where Q is the number of bits of the sampled received signal of each antenna. Hence the fronthaul load obviously grows with the number of antennas at each AP and the number of APs, and it becomes severe in large-scale deployment with a large number of antennas.

III. Hybrid signal processing

In order to reduce the fronthaul loads of the centralized processing, in this section we propose a hybrid signal processing architecture for uplink reception. Fig. 1 shows the architecture of the proposed uplink reception including APs, CPU and fronthaul. It consists of a local signal processing unit (LSPU) at each AP and a CPU that collects the output of the LSPU of all APs through fronthaul and detects the transmitted symbols of all UEs. The LSPU combines signals of N receive antennas with weight matrix $\mathbf{U}_l$ to obtain P_l outputs. The P_l dimensional output signal vector of the LSPU can be written as $\mathbf{y}_l(i) = \mathbf{U}_l \mathbf{r}_l(i)$ and it is forwarded through the fronthaul. Then, the CPU detects signals based on the following equivalent channel expression including propagation channel and local signal processing weights of each AP after collecting the fronthaul data of all APs;

$$\mathbf{y}(i) = \mathbf{H}^{(\mathrm{E})} \mathbf{s}(i) + \mathbf{z}(i), \tag{4}$$

where,

$$\mathbf{y}(i) = \begin{bmatrix} \mathbf{y}_1(i) \\ \mathbf{y}_2(i) \\ \vdots \\ \mathbf{y}_L(i) \end{bmatrix}, \mathbf{H}^{(\mathrm{E})} = \begin{bmatrix} \mathbf{U}_1 \mathbf{H}_1 \\ \mathbf{U}_2 \mathbf{H}_2 \\ \vdots \\ \mathbf{U}_L \mathbf{H}_L \end{bmatrix}, \mathbf{z}(i) = \begin{bmatrix} \mathbf{U}_1 \mathbf{n}_1(i) \\ \mathbf{U}_2 \mathbf{n}_2(i) \\ \vdots \\ \mathbf{U}_L \mathbf{n}_L(i) \end{bmatrix} \tag{5}$$

Based on the equation, the CPU can detect the signals by general MIMO detection algorithms. To keep the scalability of the CPU processing, PMMSE can be applied by rewriting the the weight vector $\mathbf{v}_k^{(\mathrm{HP})}$ corresponding to (3) as follows:

$$\mathbf{v}_k^{(\mathrm{HP})} = (\sum_{i \in \mathcal{D}_k} \mathbf{D}_k^{(\mathrm{HP})} \hat{\mathbf{h}}_i^{(\mathrm{E})} \hat{\mathbf{h}}_i^{(\mathrm{E})\mathrm{H}} \mathbf{D}_k^{(\mathrm{HP})} + \mathbf{R}_k^{(\mathrm{ZD})})^{\dagger} \mathbf{D}_k^{(\mathrm{HP})} \hat{\mathbf{h}}_k^{(\mathrm{E})}, \tag{6}$$

where $\hat{\mathbf{h}}_k^{(\mathrm{E})}$ is the estimated vector of the equivalent channel of UE k and it corresponds to the k-th column of $\mathbf{H}^{(\mathrm{E})}$. Note that the equivalent channel vector $\mathbf{h}^{(\mathrm{E})}$ can be estimated by utilizing the pilot signals and corresponding received signal parts by assuming that the received pilot signal part of the received signal processed and forwarded in the same manner as the remaining signal parts. $\mathbf{R}_k^{(\mathrm{ZD})}$ is a covariance matrix of the corresponding part of the equivalent noise vector $\mathbf{z}(i)$ after clustering and it can be written as $\mathbf{R}_k^{(\mathrm{ZD})} = \mathbf{D}_k^{(\mathrm{HP})} E[\mathbf{z}(i)\mathbf{z}^{\mathrm{H}}(i)]\mathbf{D}_k^{(\mathrm{HP})}$. The clustering matrix $\mathbf{D}_k^{(\mathrm{HP})}$ also arrays the sub-matrix $\mathbf{D}_{kl}$ diagonally as $\mathbf{D}_k$, but the size of the sub-matrix has changed from $N \times N$ to $P_l \times P_l$. It is also worth mentioning that, the clustering can be operated independently from the local signal processing weights by applying clustering based on the large-scale fading estimated from the UE measurement reports as discussed in the section II. For this purpose, the APs require to forward the reported data measured by UEs to CPU separately from the data and pilot signal part, but this amount is small, due to the slow change of the large-scale fading. The spectral efficiency of each UE can be written as $C_k^{(\mathrm{HP})} = \log_2(1 + \gamma_k^{(\mathrm{HP})})$, where $\gamma_k^{(\mathrm{HP})}$ is the post SINR of the UE calculated as follows.

$$\gamma_k^{(\mathrm{HP})} = \frac{\rho |\mathbf{v}_k^{(\mathrm{HP})} \mathbf{D}_k \hat{\mathbf{h}}_k^{(\mathrm{E})}|^2}{\rho \left(\sum_{k' \neq k}^{K} |\mathbf{v}_k^{(\mathrm{HP})} \mathbf{D}_k \mathbf{h}_{k'}^{(\mathrm{E})}|^2 + |\mathbf{v}_k^{(\mathrm{HP})} \mathbf{D}_k \tilde{\mathbf{h}}_k^{(\mathrm{E})}|^2 \right) + |\mathbf{v}_k^{(\mathrm{HP})} \mathbf{D}_k \mathbf{U}|^2}, \quad (7)$$

where $\tilde{\mathbf{h}}_k = \mathbf{h}_k - \hat{\mathbf{h}}_k$, and $\mathbf{U} = diag[\mathbf{U}_1, \cdots \mathbf{U}_l]$.

By simply setting $P_l \leq N$ for all APs, the fronthaul loads can be reduced. In order to still allow effective signal detection at the CPU, the following three schemes are considered.

A. MMSE based scheme

One possible scheme as the local processing at the AP is MMSE, which reduces interference from other UEs to extract specific selected UEs in each AP by utilizing degrees of the freedom of the antennas at AP. The MMSE weight is calculated as $\mathbf{U}_l = \hat{\mathbf{H}}_l^{(\mathrm{D})} \mathbf{R}_l^{\dagger}$, where $\hat{\mathbf{H}}_l^{(\mathrm{D})}$ is the $N \times P_l$ matrix of the estimated channel whose columns correspond to the selected UEs and $\mathbf{R}_l$ is a covariance matrix of the received signal vector $\mathbf{r}_l(i)$. It assumes that each AP select P UEs for the extraction uniformly and this corresponds to $P = P_1 = \cdots P_L$ for simplicity. P is a design parameter satisfying $N > P$ to reduce fronthaul loads; the spectral efficiency also depends on it. Concretely, AP l selects the P target UEs that have the largest $\beta_{l,k}$.

Since interference from $K - P$ UEs remains significantly, in the typical case of $N < K$, but the post processing at the CPU can alleviate it by utilizing PMMSE. Hence in this hybrid processing the interference can be cooperatively removed by AP and CPU processing.

B. EVD based schemes (LVS-EVD, GVS-EVD)

The other schemes used for local processing at the AP are EVD based schemes, utilizing the eigenvectors of $\mathbf{R}_l$ corresponding to the larger eigenvalues as the weight vectors to extract the principal components from the received signals and forward them to the CPU. Concretely, $\mathbf{R}_l$ can be written as $\mathbf{R}_l = \mathbf{E}_l \mathbf{\Lambda}_l \mathbf{E}_l^{\mathrm{H}}$ by EVD, where $\mathbf{E}_l = [\mathbf{e}_{l,1}, \cdots \mathbf{e}_{l,N}]$, and $\mathbf{\Lambda}_l = diag[\lambda_{l,1}, \cdots \lambda_{l,N}]$, respectively. $\lambda_{l,i}$ denotes the i-th largest eigenvalue of $\mathbf{R}_l$ and $\mathbf{e}_{l,i}$ is its corresponding eigenvector, i.e. $\lambda_{l,1} \geq \cdots \geq \lambda_{l,N}$. Then the weight vector at AP l can be written by selecting P_l vectors as $\mathbf{U}_l = [\mathbf{e}_{1,l}, \cdots \mathbf{e}_{l,P_l}]$.

In order to configure the number of output ports for each AP P_l, two schemes are considered. One is LVS, which utilizes a uniform number of output ports, i.e. $P = P_1 = \cdots P_L$, and P is a design parameter. In this case, each AP can select vectors by itself without sharing any information with other APs or the CPU. The other is GVS which configures P_l non-uniformly among APs and those values are adaptively determined at the CPU. To select the effective eigenvectors throughout all the APs, P_L is determined by the following procedure. At first, all the APs forward the eigenvalues $\lambda_{l,i}$ to the CPU. Next, the CPU selects the larger eigenvalues for a predetermined amount X from all the eigenvalues, i.e. $\lambda_{l,1}, \cdots \lambda_{l,N}$ for all l, without separating APs. The CPU feedback the information how many eigenvalues are selected from each AP, and that corresponds to P_l. As a result the P_l satisfies $X = \sum_{l=1}^{L} P_l$. Depending on the selected results, some AP may use all ports, i.e., $P_l = N$, or some APs may have no ports, i.e., $P_l = 0$, at an instance, but if setting X satisfying $X < LN$ the total fronthaul load of the system can be reduced effectively. It is also noteworthy that X is a predetermined parameter that can be set independently from L and N, and that it could be operated without increasing the fronthaul load linearly to the number of APs of the whole system.

Finally, we note that the constraint on the number of ports per AP, as in LVS-EVD, is well suited if independent fibers haul back the signal from the APs, while a constraint on the total number of ports is more meaningful either with a fiber fronthaul with daisy-chain configuration, or wireless fronthaul where the total spectral resources for the fronthaul are constrained.

IV. Performance Evaluation

In order to verify the effectiveness of the proposed architecture and compare the proposed algorithms including the impact of specific parameters, computer simulations have been conducted. The simulations assume a 1km x 1km target area, where all APs and UEs are randomly distributed. Basic parameters are listed in Table I. The number of APs L and number of antennas at AP N, is set to 64 and 4, respectively. The large-scale fading generated in the simulation is given as follows based on [8]:

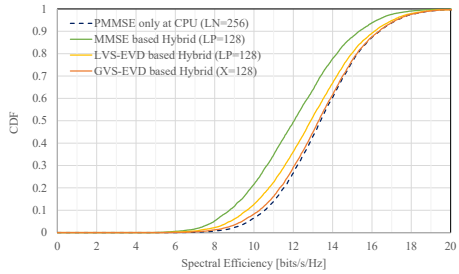
$$\beta_{l,k} = g_0 - 10\gamma \log_{10}\left(\frac{d_{l,k}}{d_0}\right) + \frac{\sigma_w^2}{\sqrt{2}}\left(w_l^{\mathrm{AP}} + w_k^{\mathrm{UE}}\right), \quad (8)$$

where $d_{l,k}$ is the distance between AP l and UE k. w_l^{AP} and w_k^{UE} are normalized shadow fading of AP l and UE k, respectively, and σ_w^2 is the variance. Although shadowing is related to the link, and not separately of the AP and UE, splitting the total link shadowing into two contributions following [5], [12] is executed to (approximately) consider

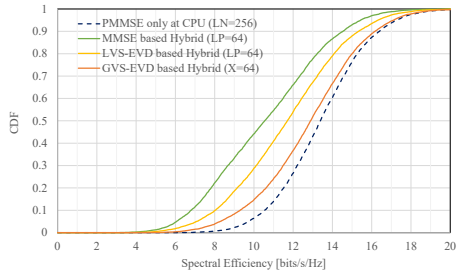
TABLE I
BASIC SIMULATION PARAMETERS

Number of APs (L)	64
Number of antennas of AP (N)	4
Number of spatial multiplexing UEs (K)	16, 64
Size of the cluster for each UE (Z)	8, 32
Carrier frequency	3.5 GHz
Bandwidth	100 MHz
Transmission power ($\bar{P}$)	23 dBm
Noise power	-87 dBm
Fading	Rician
Rician K-factor in dB	$13 - 0.03 \cdot$ Distance
Median channel gain at d_0 (g_0)	-43.3 dB
Path loss exponent (γ)	2
Azimuth angular standard deviation (σ_ϕ)	20°
Channel estimation	LMMSE

the shadowing correlation between different UEs and APs, respectively. The phase of the line-of-sight component can be determined by geometrical considerations. For the channel estimation, orthogonal pilot symbols are assumed to be allocated to UEs without any contamination and 2 symbols are allocated for each UE for LMMSE channel estimation. The spectral efficiency is obtained by calculating (7) for each channel realizations with random drops of APs and UEs.



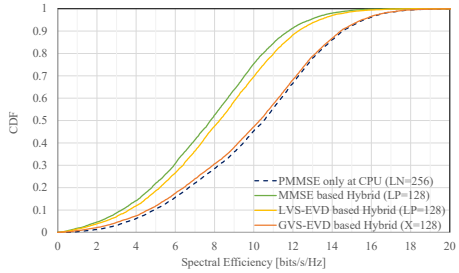
(a) Fronthaul load reduction rate = 0.5



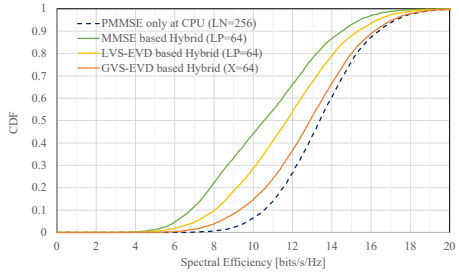
(b) Fronthaul load reduction rate = 0.75

Fig. 2. Comparison of proposed hybrid processing schemes and PMMSE when K=16, Z=32

Fig. 2 shows the CDF of the spectral efficiency of each multiplexed UE when utilizing the hybrid processing schemes with MMSE, LVS-EVD, and GVS-EVD, respectively, when setting the number of multiplexed UEs as K=16, the size of the cluster for each UE as Z=32, and the number of antennas at AP N=4. As a benchmark-1, the spectral efficiency of PMMSE only at the CPU, using full fronthaul loads when setting N=4 is shown in the same figure. The fronthaul load reduction rate by the hybrid processing scheme to the benchmark-1 is set to 0.5 and 0.75 for the results shown in (a) and (b), respectively. The rate 0.5 corresponds to configuration P=2 for MMSE and LVS-EVD, and X=128 for GVS-EVD, and the rate 0.75 corresponds to configure P=1 and X=64, respectively. In addition, PMMSE only at CPU with setting the same number of antennas N to the output ports of the hybrid processing ($N = P$) are also shown as the benchmark-2. For both the reduction rate, the remarkable degradation from the benchmark-1 due to the fronthaul load reduction can be observed for MMSE and LVS-EVD, whereas the improvement from the benchmark-2 can be observed. It can be said that, EVD based schemes are more robust than MMSE, and GVS-EVD can obtain close performance to benchmark-1 even the reduction rate is higher.



(a) Fronthaul load reduction rate = 0.5



(b) Fronthaul load reduction rate = 0.75

Fig. 3. Comparison of the proposed hybrid processing schemes and PMMSE when K=16, Z=8

Fig. 3 shows the results when reducing the cluster size to Z=8. Comparing to Z=32, the spectral efficiency itself degrades as a whole because the capability to suppress the interference decreases, but it can reduce the computational complexity more than Z=32. When the rate is 0.5, as shown in (a), GVS-EVD can obtain the closest performance to the benchmark-1. When the rate is 0.75, severe degradation from benchmark-1 is observed for LVS-EVD and MMSE-IRC, whereas the improvement from benchmark-2 can be observed a little. Only the GVS-EVD can still keep the degradation smaller. Hence it can be said that GVS-EVD is robust to setting the cluster size smaller. This is also important feature especially for large scale networks to operate with realistic computational complexity.

Fig. 4 shows the relationships of the degradation of the median spectral efficiency from the benchmark-1 for GVS-EVD with the parameter X for Z=8 and 32, respectively.

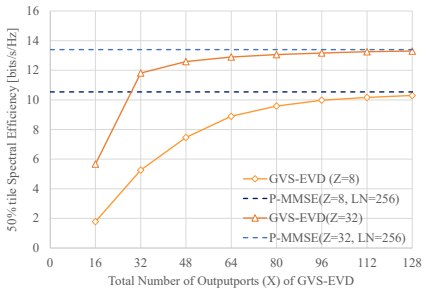
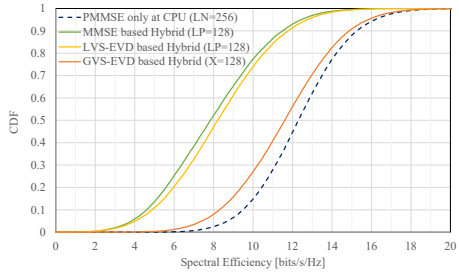
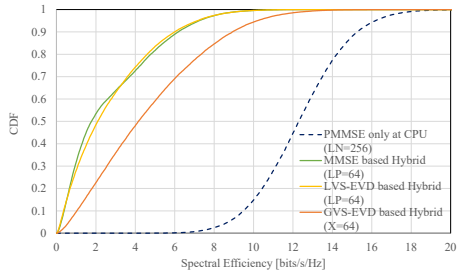


Fig. 4. Relationships between total number of output ports of GVS-EVD and median spectral efficiency

Setting Z larger, the degradation remains smaller even if X is smaller, as the fronthaul load reduction rate, $1 - X/LN$, increases. When Z=32, the degradation is kept within 10% up to the reduction rate 0.875, corresponding X=32, and the rate 0.6875, corresponding to X=80, when Z=8.



(a) Fronthaul load reduction rate = 0.5



(b) Fronthaul load reduction rate = 0.75

Fig. 5. Comparison of the proposed hybrid processing schemes and PMMSE when K=64, Z=32

Fig. 5 shows the results when setting K=64 and Z=32. Compared to the case when K=16, shown in Fig.2, the spectral efficiency of each UE is degraded as a whole and degradation of the hybrid processing from the benchmark-1 is not negligible even with the GVS-EVD. However, it can be said that the closer performance can be obtained compared to the other schemes when the reduction rate is 0.5, and this is also robust to increase of the number of UEs.

Throughout these evaluations, the effectiveness of the hybrid processing architecture to reduce the fronthaul load could be verified. However, when employing MMSE and LVS-EVD, which reduce the fronthaul load uniformly in all APs, the improvement from the benchmark-2 is limited and remarkable degradation from the benchmark-1 is observed, especially for the severe situations such as small cluster size or large number of UEs. In contrast, GVS-EVD was found to be able to maintain reasonable performance to some extent even under the severe situations. This superiority could be explained because it can flexibly control the data amount from each AP to the CPU under each situation according to the importance of the AP that depends on distribution of APs and UEs.

V. Conclusion

This paper has proposed a hybrid signal processing architecture for scalable uplink cell-free massive MIMO to reduce fronthaul loads for a scalable centralized processing architecture. In the architecture, local signal processing, which compresses the fronthaul loads, can be cooperatively operated with PMMSE at the CPU while keeping the scalability of the computational complexity to detect each UE. For this architecture we have considered MMSE, namely LVS-EVD and GVS-EVD, as the local processing. Computer simulation results show the architecture can effectively reduce the fronthaul loads while retaining good spectral efficiency. It also shows that GVS-EVD can obtain the best performance among the three schemes and is robust to increase the fronthaul reduction rate, number of multiplexed UEs, and smaller cluster size operation. Further evaluations, such as the effect of the quantization noise for low resolution ADCs, that of the channel estimation error caused by the pilot contamination, and number of antennas, will be a future work.

References

[1] 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, Mar. 2017.

[2] Ö. T. Demir, E. Björnson and L. Sanguinetti, Foundations of User-Centric Cell-Free Massive MIMO," in *Foundations and Trends® in Signal Processing*, vol. 14, no. 3–4, pp. 162–472, 2021.

[3] H. Tataria, M. Shafi, A.F. Molisch, M. Dohler, H. Sjöland, and F. Tufvesson, "6G Wireless Systems: Vision, Requirements, Challenges, Insights, and Opportunities," *Proc. IEEE*, vol. 109, no. 7, pp. 1166-1199, 2021.

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

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

[6] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, and M. Debbah, "Cell-Free Massive MIMO with Limited Backhaul," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018.

[7] J. Zhang, J. Zhang, D. W. K. Ng, S. Jin, and B. Ai, "Improving Sum-Rate of Cell-Free Massive MIMO with Expanded Compute-and-Forward," *IEEE Trans. Signal Process.*, vol. 70, pp. 202-2015, Nov. 2021.

[8] 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.

[9] R. V. Rompaey, and M. Moonen, "Scalable and Distributed MMSE Algorithms for Uplink Receive Combining in Cell Free Massive MIMO Systems," in *Proc. IEEE Int. Conf. Accoust. Speech Signal Process. (ICASSP)*, June 2021.

[10] S. Buzzi, and C. D'Andrea, "Cell-free massive MIMO: User-centric approach," *IEEE Commun. Lett.*, vol. 6, no. 6, pp. 706-709, 2017.

[11] F. R. Palou, G. Femenias, A. G. Armada, and A. Pérez-Neira,"Clustered Cell-Free Massive MIMO," in *Proc. IEEE Globecom Workshops (GC Wkshps 2018)*, Dec. 2018.

[12] Ö. Özdogan, E. Björnson and J. Zhang, "Performance of cell-free massive MIMO with Rician fading and phase shifts," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5299–5315, Nov. 2019.