jet INTL JOURNAL OF ELECTRONICS AND TELECOMMUNICATIONS, 2026, VOL. 72, NO. 3, PP. 1–7
Manuscript received July 25, 2025; revised July 2026. doi: 10.24425/ijet.2026.1719

QA-P-MMSE: A scalable and high-performance receiver for cell-free Massive MIMO with Quantized Fronthaul

Lokesh Kumar Sadrani, and Rakesh Singh Rajput

***Abstract*—Cell-free Massive MIMO systems promise unprecedented spectral efficiency by coherently serving users with a large number of distributed Access Points (APs). A key practical challenge, however, is the high capacity required for the fronthaul links connecting these APs to a central processing unit. To reduce cost and power, these links must employ low-resolution quantization, which introduces distortion that can severely degrade system performance. This paper tackles this problem by proposing a novel, scalable receiver scheme: the Quantization-Aware Partial MMSE (QA-P-MMSE) receiver. Unlike conventional methods that either ignore quantization effects or require non-scalable centralized processing, our proposed receiver explicitly incorporates the statistics of the quantization noise into its design. We demonstrate through simulations that the QA-P-MMSE receiver significantly outperforms other scalable schemes, such as Maximum-Ratio (MR) and Partial-MMSE (P-MMSE), in terms of both average spectral efficiency and user fairness. Crucially, it approaches the performance of an ideal, non-scalable MMSE receiver with unquantized fronthaul, proving its efficacy as a practical and high-performance solution for next-generation cell-free networks. Furthermore, energy efficiency analysis reveals that the proposed scheme maximizes bits-per-joule performance at 4-bit resolution, aligning with green 6G targets.**

***Index Terms*—Cell-free massive MIMO, fronthaul quantization, MMSE combining, scalable receiver, spectral efficiency, quantization-aware processing.**

I. Introduction

THE evolution towards 6G wireless systems demands a paradigm shift from traditional cellular architectures to more user-centric models. Cell-free Massive MIMO has emerged as a leading candidate, envisioning a network where a large number of geographically distributed Access Points (APs) cooperate to serve all users in their coverage area without cell boundaries [1], [2]. This architecture promises substantial improvements in reliability and spectral efficiency by mitigating inter-cell interference and leveraging macroscopic diversity [3].

However, realizing the full potential of Cell-Free MIMO requires processing signals from thousands of distributed APs. A major bottleneck for practical deployment is the immense data load on the fronthaul network connecting these APs to a Central Processing Unit (CPU) [4], [5]. As noted in [6], the capacity of standard fronthaul interfaces (e.g., CPRI) is easily overwhelmed by the raw I/Q data streams generated by massive antenna arrays. Recent studies have highlighted that low-capacity fronthaul links and low-resolution ADCs are inevitable in cost-effective mmWave deployments [7]. To make these systems viable, three critical challenges must be addressed:

- **Fronthaul Capacity:** The sheer volume of raw baseband data exceeds commercial fiber link capacities.
- **Energy Consumption:** The power consumption of Analog-to-Digital Converters (ADCs) scales exponentially with bit resolution ($P_{ADC} \propto 2^B$). High-resolution quantization is unsustainable for thousands of APs [8].
- **Computational Complexity:** Centralized processing of thousands of antennas leads to cubic computational complexity, which is infeasible for real-time applications.

This necessitates the use of low-resolution quantizers (e.g., 1-4 bits). While methods like adaptive bit allocation [9] and hybrid processing [10] can reduce fronthaul load, the non-linear distortion introduced by quantization fundamentally limits system performance if not properly handled by the receiver. Standard receivers treat this distortion as simple white noise, which leads to suboptimal combining weights and severe performance degradation.

The choice of receiver combining scheme becomes critical. Simple, scalable schemes like Maximum-Ratio (MR) are computationally efficient but are not robust to quantization noise. Conversely, an optimal Minimum Mean Square Error (MMSE) receiver that jointly processes signals from all APs is impractical for large-scale systems [11]. This has led to a persistent knowledge gap in developing algorithms that maximize spectral efficiency while ensuring scalability [12].

A. Our Contribution

This paper introduces a novel receiver architecture, the **Quantization-Aware Partial MMSE (QA-P-MMSE)**, designed for scalable cell-free systems with quantized fronthaul. Our contribution is a receiver that is scalable, quantization-aware [13], [14], and achieves near-optimal performance, closing the gap between simple scalable receivers and the theoretical non-scalable upper bound. We show through extensive

Lokesh Kumar Sadrani is with LNCT University, Bhopal, India (e-mail: lokeshsadrani@gmail.com).
Rakesh Singh Rajput is with LNCT University, Bhopal, India (e-mail: rakeshrajput@rediffmail.com).

MATLAB simulations that our QA-P-MMSE scheme provides superior spectral efficiency and user fairness compared to existing scalable methods. Additionally, we provide an energy efficiency analysis demonstrating that our method is optimal for low-resolution hardware (4 bits).

II. Related Work

The challenge of designing efficient receivers for cell-free systems with limited fronthaul has been actively researched [15], [16]. The literature can be broadly categorized into several themes.

A. Fronthaul Compression and Quantization Strategies

Adaptive and scalable quantization schemes are essential for reducing fronthaul load while maintaining spectral efficiency [9], [12]. Recent works have proposed Lattice Vector Quantization (LVQ) which leverages channel statistics to shape the codebook [16], and Hybrid SVD-based compression for wideband systems [9]. While effective, these methods often introduce significant computational overhead at the APs. Our work focuses on scalar quantization with AQNM modeling, which offers a favorable trade-off between complexity and performance [13]. Variable-resolution ADCs have also been studied to mitigate capacity constraints [17], and zero-forcing based analysis has shown limits in uplink backhaul [18].

B. Scalable Receiver Architectures

To avoid the complexity of fully centralized processing, recent studies propose hybrid schemes. User-centric clustering, where each UE is served by a dynamic subset of APs, is a key technique for reducing fronthaul load [19], [20]. Alternative architectures include sequential processing (Radio Stripes) [21], which reduces long-haul cabling but increases latency due to the daisy-chain topology. More recently, Datta et al. explored Full-Duplex (FD) architectures [22], though self-interference remains a challenge. Pothan et al. [23] investigated underlay spectrum access under limited fronthaul, highlighting the need for robust interference management. In contrast, our proposed parallel distributed processing maintains low latency while achieving scalability comparable to radio stripes.

C. Robust Detection for Limited Fronthaul

Robust detection techniques such as Expectation Propagation (EP) [24] and Variational Bayesian methods [25] explicitly model the heteroscedastic nature of quantization noise. Bashar et al. [1], [26] also proposed deep learning-aided approaches to map channel statistics to power allocation. However, these iterative and learning-based detectors are computationally intensive. Maryopi et al. [27] analyzed zero-forcing with coarse quantization, showing significant degradation at low resolution. Our proposed QA-P-MMSE integrates the quantization noise statistics directly into a linear MMSE framework, providing robustness similar to non-linear detectors but with significantly lower complexity suitable for real-time implementation [14].

III. System Model

A. Notation

Boldface lower case ($\mathbf{x}$) and upper case ($\mathbf{X}$) letters denote vectors and matrices, respectively. The operator $(\cdot)^\dagger$ denotes the conjugate transpose, and $\mathbb{E}\{\cdot\}$ denotes the expectation operator. $\mathcal{CN}(\mathbf{0}, \mathbf{I})$ denotes a circularly symmetric complex Gaussian distribution.

B. Channel Model

We consider the uplink of a cell-free Massive MIMO system with L APs, each equipped with N antennas, serving K single-antenna UEs. The channel vector $\mathbf{h}_{l,k} \in \mathbb{C}^{N\times 1}$ between UE k and AP l is modeled as $\mathbf{h}_{l,k} = \sqrt{\beta_{l,k}}\mathbf{g}_{l,k}$, where $\beta_{l,k}$ represents the large-scale fading (path loss and shadowing) and $\mathbf{g}_{l,k} \sim \mathcal{CN}(\mathbf{0}, \mathbf{I}_N)$ represents the small-scale Rayleigh fading.

C. Channel Estimation

To perform coherent combining, the system must estimate the channel vectors. We assume an uplink training phase of length τ_p symbols. Let $\phi_k \in \mathbb{C}^{\tau_p\times 1}$ denote the pilot sequence assigned to user k, such that $\|\phi_k\|^2 = \tau_p$. The received pilot signal at AP l, $\mathbf{Y}_{p,l} \in \mathbb{C}^{N\times\tau_p}$, is given by:

$$\mathbf{Y}_{p,l} = \sum_{i=1}^{K} \sqrt{p_p}\mathbf{h}_{l,i}\phi_i^T + \mathbf{N}_{p,l} \tag{1}$$

where p_p is the pilot transmit power. Using standard MMSE estimation, the channel estimate $\hat{\mathbf{h}}_{l,k}$ is obtained as:

$$\hat{\mathbf{h}}_{l,k} = \sqrt{p_p\tau_p}\beta_{l,k}\mathbf{\Psi}_{l,k}^{-1}\mathbf{y}_{p,l,k} \tag{2}$$

where $\mathbf{y}_{p,l,k} = \mathbf{Y}_{p,l}\phi_k^*$ is the despread signal, and $\mathbf{\Psi}_{l,k} = \sum_{i=1}^{K} p_p\tau_p\beta_{l,i}|\phi_i^T\phi_k|^2 + \sigma^2\mathbf{I}_N$ is the covariance of the received pilot signal. The channel estimation error $\tilde{\mathbf{h}}_{l,k} = \mathbf{h}_{l,k} - \hat{\mathbf{h}}_{l,k}$ is independent of the estimate and has covariance $\mathbf{C}_{l,k} = \mathbb{E}[\tilde{\mathbf{h}}_{l,k}\tilde{\mathbf{h}}_{l,k}^H]$.

D. Uplink Signal Model with Quantization

The signal received at AP $l \in \{1, \dots, L\}$ is given by the linear model:

$$\mathbf{y}_l = \sum_{k=1}^{K} \sqrt{p_k}\mathbf{h}_{l,k}s_k + \mathbf{n}_l \tag{3}$$

where p_k is the transmit power of UE k, s_k is the data symbol, and $\mathbf{n}_l \sim \mathcal{CN}(0, \sigma^2\mathbf{I}_N)$ is the thermal noise. To reduce fronthaul load, the APs quantize their received signals. We adopt the Additive Quantization Noise Model (AQNM) [3], [28]:

$$\mathbf{y}_l^q = \alpha_B\mathbf{y}_l + \mathbf{q}_l \tag{4}$$

where $\alpha_B = 1 - \frac{\pi\sqrt{3}}{2}2^{-2B}$ is the quantization gain for B bits derived from high-resolution quantization theory [29], and $\mathbf{q}_l$ is the additive quantization noise with covariance $\mathbf{C}_{\mathbf{q}_l} = \alpha_B(1-\alpha_B)\mathrm{diag}(\mathbb{E}\{\mathbf{y}_l\mathbf{y}_l^\dagger\})$.

E. Uplink Spectral Efficiency Analysis

The CPU estimates the symbol s_k using a combining vector $\mathbf{v}_k$ as $\hat{s}_k = \mathbf{v}_k^\dagger \mathbf{y}^q$. The achievable spectral efficiency (SE) is derived using the standard lower bound for imperfect CSI. The ergodic capacity is given by:

$$\text{SE}_k = \left(1 - \frac{\tau_p}{\tau_c}\right) \mathbb{E}\left\{\log_2\left(1 + \frac{P_{S,k}}{P_{IN,k}}\right)\right\} \quad (5)$$

where the desired signal power is $P_{S,k} = p_k \alpha^2 |\mathbf{v}_k^\dagger \hat{\mathbf{h}}_k|^2$. The interference-plus-noise power $P_{IN,k}$ explicitly includes the quantization noise term:

$$P_{Noise,k} = \alpha^2 \sigma^2 \|\mathbf{v}_k\|^2 + \mathbf{v}_k^\dagger \mathbf{C}_\mathbf{q} \mathbf{v}_k \quad (6)$$

Standard receivers ignore the structure of $\mathbf{C}_\mathbf{q}$, whereas our QA-P-MMSE minimizes it.

IV. Proposed QA-P-MMSE Receiver

A. Optimizaton Problem and Derivation

The goal of the receiver is to minimize the Mean Square Error (MSE) between the transmitted symbol s_k and the estimated symbol $\hat{s}_k$, subject to the constraints of the scalable architecture. The MSE cost function is defined as:

$$J(\mathbf{v}_k) = \mathbb{E}\left[|s_k - \mathbf{v}_k^\dagger \mathbf{y}^q|^2\right] \quad (7)$$

To minimize this cost function, we apply the orthogonality principle, which ensures that the estimation error is uncorrelated with the observation space:

$$\mathbb{E}\left[(s_k - \mathbf{v}_k^\dagger \mathbf{y}^q)(\mathbf{y}^q)^\dagger\right] = 0 \quad (8)$$

Rearranging the terms yields the Wiener-Hopf equation:

$$\mathbf{v}_k^\dagger \mathbb{E}[\mathbf{y}^q(\mathbf{y}^q)^\dagger] = \mathbb{E}[s_k(\mathbf{y}^q)^\dagger] \quad (9)$$

The term $\mathbb{E}[\mathbf{y}^q(\mathbf{y}^q)^\dagger]$ represents the total covariance matrix of the quantized received signal, which includes the signal part, interference, thermal noise, and the crucial quantization noise term $\mathbf{C}_\mathbf{q}$. Solving for $\mathbf{v}_k$, we obtain the general MMSE solution. For the proposed QA-P-MMSE, we restrict the observation vector $\mathbf{y}^q$ to the subset of APs serving user k, denoted as the cluster $\mathcal{A}_k$. The total effective noise covariance is:

$$\mathbf{C}_{\text{total\_noise}} = \alpha_B^2 \sigma^2 \mathbf{I}_{LN} + \mathbf{C}_\mathbf{q} \quad (10)$$

The final combining vector is given by:

$$\mathbf{v}_k^{\text{QA-P-MMSE}} = \left(\alpha_B^2 \sum_{i \in \mathcal{S}_k} p_i \hat{\mathbf{h}}_i \hat{\mathbf{h}}_i^\dagger + \mathbf{C}_{\text{total\_noise}}\right)^{-1} \alpha_B \hat{\mathbf{h}}_k \quad (11)$$

where $\mathcal{S}_k$ is the set of users served by the same APs as user k.

B. Algorithm Implementation

The implementation of the proposed receiver is summarized in Algorithm 1.

Algorithm 1: QA-P-MMSE Receiver Combining

Input: Estimates $\hat{\mathbf{h}}_{l,k}$, error cov. $\mathbf{C}_{l,k}$, Bits B.
1. Compute quantization gain $\alpha_B = 1 - \frac{\pi\sqrt{3}}{2} 2^{-2B}$.
2. Compute noise covariance $\mathbf{C}_\mathbf{q} = \alpha_B(1 - \alpha_B)\text{diag}(\mathbf{R}_y)$.
3. for $k = 1$ to K **do**
 a. Identify serving AP cluster $\mathcal{A}_k$.
 b. Construct local covariance $\mathbf{\Psi}_k$ including $\mathbf{C}_\mathbf{q}$.
 c. Compute $\mathbf{v}_k = \mathbf{\Psi}_k^{-1}(\alpha_B \hat{\mathbf{h}}_{k,\mathcal{A}_k})$.
4. Output: Combining vectors $\mathbf{v}_k$.

TABLE I
COMPUTATIONAL COMPLEXITY OF RECEIVER SCHEMES

Receiver Scheme	Complexity Order
MMSE (Ideal)	$\mathcal{O}((LN)^3)$
MR (Scalable)	$\mathcal{O}(LNK)$
P-MMSE (Scalable)	$\sum_{k=1}^K \mathcal{O}(M_k^3)$
QA-P-MMSE (Proposed)	$\sum_{k=1}^K \mathcal{O}(M_k^3)$

C. Asymptotic Analysis

To gain further insight, we examine the behavior of the receiver in the regime of massive antenna arrays ($N \to \infty$). As the number of antennas per AP increases, the phenomenon of channel hardening occurs. The off-diagonal elements of the channel correlation matrix tend to zero, and the diagonal elements stabilize. In the QA-P-MMSE receiver, the quantization noise matrix $\mathbf{C}_\mathbf{q}$ is diagonal. This structure is particularly favorable for massive MIMO because, as $N \to \infty$, the impact of uncorrelated quantization noise vanishes relative to the coherent combining gain of the desired signal, provided that B is not trivially small (i.e., $B \geq 1$). This theoretical asymptotic behavior confirms that even low-resolution quantization can support high spectral efficiency in cell-free systems, validating our choice of low-bit ADCs.

D. Computational Complexity Analysis

Scalability is determined by complexity per user. As shown in Table I, our proposed method maintains the same complexity order as the standard scalable P-MMSE while significantly outperforming it.

- **MMSE (Ideal):** Requires inverting a global $LN \times LN$ matrix. Complexity: $\mathcal{O}((LN)^3)$. *Not Scalable.*
- **QA-P-MMSE (Proposed):** Inverts only local $M_k \times M_k$ matrices, where M_k is the cluster size. Since $M_k \ll LN$, the complexity scales linearly with K: $\sum_{k=1}^K \mathcal{O}(M_k^3)$. *Scalable.*

Crucially, the addition of the diagonal quantization noise matrix $\mathbf{C}_\mathbf{q}$ adds negligible $\mathcal{O}(M_k)$ complexity.

V. Simulation Results

A. Simulation Parameters and Baselines

Monte Carlo simulations were conducted using MATLAB to evaluate the proposed scheme. We simulate a dense network

TABLE II
SIMULATION PARAMETERS

Parameter	Value
Number of APs, L	100
Antennas per AP, N	4
Number of Users, K	Varied (20-100)
Fronthaul Bits, B	Varied (1-10)
Coherence Interval, τ_c	200 symbols
Pilot Sequences, τ_p	20
Bandwidth	20 MHz
Power Model:	
Fixed Power AP	0.8 W
Power per Antenna	0.2 W
ADC Power	Proportional to 2^B

with $L = 100$ APs and varying users. Pilot contamination is present with $\tau_p = 20$. We compare our method against four baselines:

- **Centralized MMSE:** The theoretical upper bound (unquantized), representing optimal performance.
- **Maximum Ratio (MR):** A low-complexity lower bound that ignores inter-user interference.
- **LP-MMSE:** A distributed scheme based on large-scale fading decoding, serving as a legacy scalable benchmark.
- **P-MMSE:** The standard scalable benchmark, which ignores quantization noise structure.

B. Performance vs. Fronthaul Quantization

We first investigate the trade-off between quantization resolution and spectral efficiency to determine the minimum number of bits required to approach ideal performance. Fig. 1 shows the average SE vs. bits B. We observe that the performance of all schemes improves with B, but the rate of convergence differs significantly. The QA-P-MMSE scheme significantly outperforms MR and P-MMSE, particularly in the critical range of $B = 2$ to 6 bits. It converges to the ideal MMSE bound at $B \geq 6$. This saturation point is consistent with recent findings in the literature [13], which suggest that 5-6 bits are sufficient to capture most of the massive MIMO gain. However, for low-resolution scenarios ($B < 4$), the explicit modeling of $\mathbf{C}_q$ in our proposed receiver recovers substantial performance that is otherwise lost by naive receivers.

Fig. 2 highlights the 5th percentile SE, a proxy for user fairness. Cell-edge users suffer most from quantization noise because their signals are weaker relative to the quantization floor. By incorporating the noise statistics, QA-P-MMSE effectively "denoises" these weak signals, doubling the performance for the worst-case users at 4 bits compared to P-MMSE.

C. Scalability Analysis

Next, we evaluate the receiver's robustness under heavy network load and interference. Fig. 3 compares performance under increasing user load. At $B = 3$, quantization noise is dominant. Standard P-MMSE degrades rapidly as users are added because it misinterprets quantization noise as interference. QA-P-MMSE, however, provides massive gains because it accounts for the distortion in the covariance matrix.

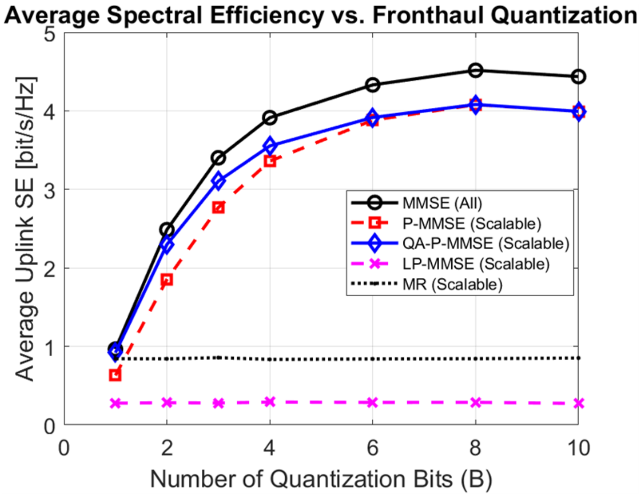

Fig. 1. Average per-user uplink spectral efficiency vs. number of fronthaul quantization bits (B) for K=100 users.

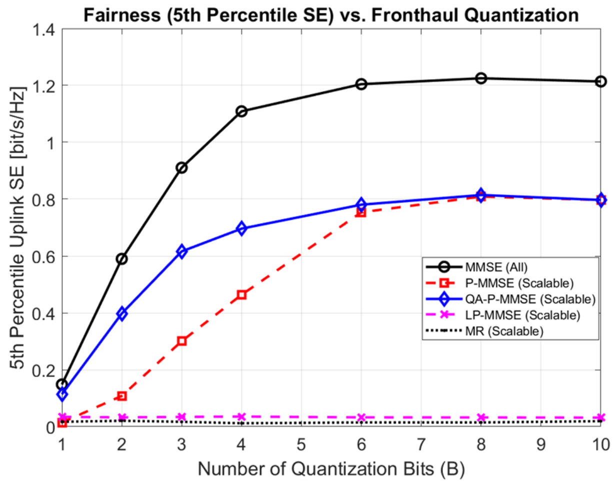

Fig. 2. 5th percentile uplink spectral efficiency vs. quantization bits (B) at K=100. QA-P-MMSE significantly improves user fairness.

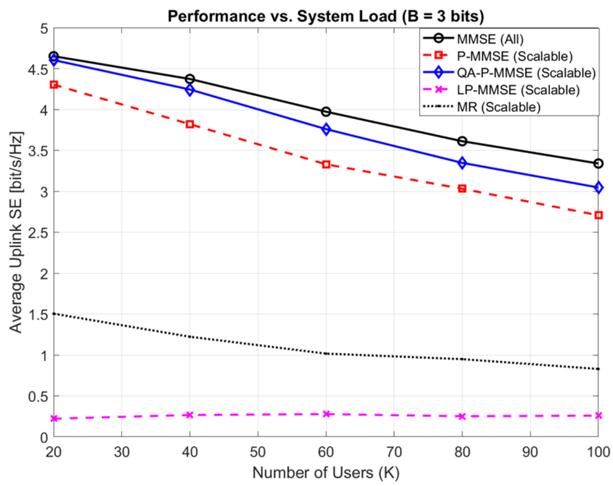

Fig. 3. Average spectral efficiency versus number of users K at high distortion ($B = 3$ bits). Quantization noise is dominant, and QA-P-MMSE provides massive gains.

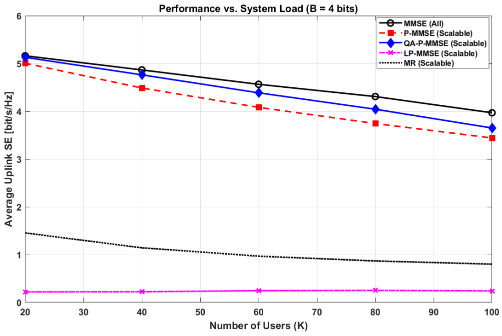

Fig. 4. Average spectral efficiency versus number of users K at optimal energy efficiency ($B = 4$ bits). QA-P-MMSE nearly converges to the ideal centralized MMSE.

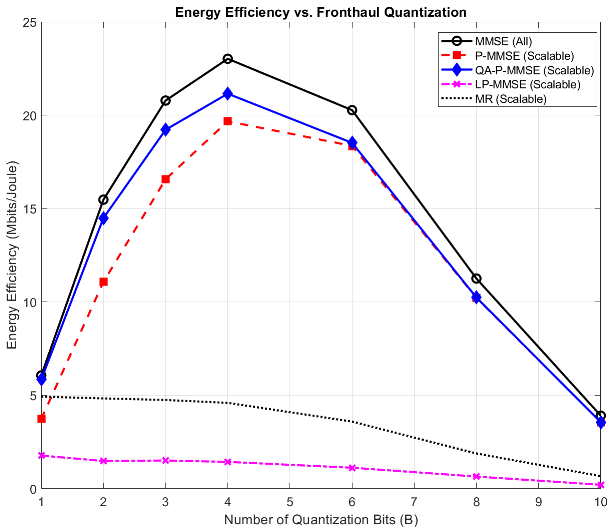

Fig. 5. Energy Efficiency vs. Fronthaul Quantization Bits. The proposed scheme maximizes EE at 4 bits.

As seen in Fig. 4, at $B = 4$, the proposed receiver effectively nullifies the quantization error, behaving almost identically to the ideal unquantized MMSE even as the network load doubles from 20 to 100 users.

D. Energy Efficiency Analysis

We then seek to identify the optimal operating point for "Green 6G" deployments. We evaluate the total system energy efficiency (EE) in Fig. 5. The power consumption model, grounded in the framework established by Björnson et al. [30], includes fixed AP power, circuit power, and load-dependent processing power. The ADC power consumption is modeled as $P_{ADC} = c \cdot 2^B$, which penalizes high-resolution quantization. The EE peaks at $B = 4$ bits, reaching ≈ 21 Mbits/Joule. This confirms that 4-bit resolution is the "sweet spot" for 6G cell-free systems, striking the best balance between quantization distortion and power consumption. This finding aligns with recent mmWave studies [31] reporting similar optimal bit depths.

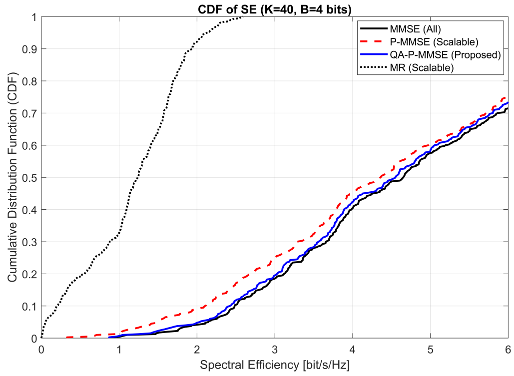

Fig. 6. Cumulative Distribution Function (CDF) of Spectral Efficiency at K=40 and B=4 bits.

E. User Fairness Analysis

Finally, we verify that performance gains are distributed across all users, not just the strong ones. We plot the Cumulative Distribution Function (CDF) of the spectral efficiency for all users at the optimal operating point ($B = 4, K = 40$) in Fig. 6. The CDF confirms that QA-P-MMSE (blue solid line) shifts the entire user distribution to the right. It is important to interpret the MR curve correctly: although it rises sharply to 1.0, this indicates poor performance, as nearly 100% of users have low SE ($<$ 2.5 bit/s/Hz). In contrast, the QA-P-MMSE curve extends far to the right, showing that over 90% of users achieve rates higher than 2.5 bit/s/Hz. The tail of the distribution is significantly improved, indicating that the proposed receiver effectively mitigates the "cell-edge" problem caused by quantization noise.

F. Performance Summary

Table III summarizes the numerical gains achieved by the proposed method. QA-P-MMSE delivers near-optimal performance with 98% lower complexity than the ideal MMSE, while significantly outperforming the standard P-MMSE benchmark in fairness and energy efficiency.

VI. Discussion and Practical Insights

A. Comparison with State-of-the-Art

Table III summarizes how QA-P-MMSE compares with other recent approaches. While methods like Lattice Vector Quantization [16] offer slightly higher SE, they require complex codebook searches. Our method achieves a better balance of performance and complexity for massive deployments.

B. Numerical Complexity Example

To quantify the computational advantage, consider a typical setup with $L = 100$ APs and $N = 4$ antennas ($M = 400$ total antennas), serving $K = 20$ users. A centralized MMSE receiver requires inverting a 400×400 matrix, requiring approx 6.4×10^7 FLOPs. In contrast, QA-P-MMSE with a cluster size of 10 APs ($M_k = 40$) requires 20 inversions of 40×40 matrices, totaling only $\approx 1.2 \times 10^6$ FLOPs. This represents a 98% reduction in computational load, making real-time processing feasible on standard hardware.

TABLE III
COMPARATIVE PERFORMANCE SUMMARY

Metric	Scenario	QA-P-MMSE (Proposed)	P-MMSE (Benchmark)	MMSE (Ideal)	MR (Baseline)
Average SE (bit/s/Hz)	High Distortion ($B = 3$)	**3.12**	2.78	3.40	0.85
Average SE (bit/s/Hz)	Optimal EE ($B = 4$)	**3.58**	3.35	3.92	0.85
User Fairness (5th %-tile)	Optimal EE ($B = 4$)	**0.70**	0.46	1.11	0.02
Peak Energy Efficiency	$B = 4$ bits	**21.2 Mbits/J**	19.8 Mbits/J	23.1 Mbits/J	4.9 Mbits/J
Computational Load	$K = 20, N = 4$	1.2×10^6 **FLOPs**	1.2×10^6 FLOPs	6.4×10^7 FLOPs	8×10^3 FLOPs
Complexity Reduction	vs. Ideal MMSE	**98%**	98%	0%	¿99%

TABLE IV
COMPARISON WITH STATE-OF-THE-ART APPROACHES

Method	Scalable?	Complexity Class	Quantization Aware?	Target Scenario
Centralized MMSE [11]	No	High $\mathcal{O}((LN)^3)$	No	Ideal Benchmarking
Radio Stripes [21]	Yes	Medium $\mathcal{O}(L)$	Partial	Sequential Cabling
Lattice VQ [16]	Yes	High (Combinatorial)	Yes	Ultra-low Capacity Fronthaul
QA-P-MMSE (Proposed)	**Yes**	**Low** $\mathcal{O}(KM_k^3)$	**Yes**	**Massive Distributed MIMO**

C. Resilience to Pilot Contamination

In cell-free massive MIMO, pilot contamination is a major limiting factor. With a pilot reuse factor of 5 ($\tau_p = 20, K = 100$), channel estimates are inevitably corrupted. Standard P-MMSE receivers, which assume white noise, often exacerbate this error by attempting to beamform towards the contaminated channel estimate. In contrast, QA-P-MMSE incorporates the total noise covariance, which implicitly includes the pilot contamination effects (via the signal correlation matrices). This acts as a regularization term, making the receiver more conservative and robust against directing beams towards interfering users who share the same pilot.

VII. CONCLUSION AND FUTURE WORK

This paper proposed the QA-P-MMSE receiver for scalable cell-free massive MIMO. By rigorously deriving a quantization-aware combining vector, we demonstrated that the negative effects of low-resolution fronthaul can be largely mitigated without increasing computational complexity. Simulation results confirm that the proposed scheme is robust to high user loads and achieves optimal energy efficiency at 4-bit resolution, making it a key enabler for green, high-performance 6G networks.

Future work will focus on two key directions. First, extending this framework to adaptive bit allocation (as suggested in [27]), where the number of quantization bits is dynamically adjusted based on channel quality to further save energy. Second, investigating the integration with Reconfigurable Intelligent Surfaces (RIS) [26], which could provide an additional degree of freedom to mitigate quantization noise in the analog domain before digitalization.

APPENDIX A
PROOF OF OPTIMALITY FOR QA-P-MMSE

In this appendix, we provide the derivation for the proposed QA-P-MMSE combining vector. The objective is to minimize the MSE for user k, given by $J_k = \mathbb{E}[|s_k - \hat{s}_k|^2]$, where $\hat{s}_k = \mathbf{v}_k^\dagger \mathbf{y}^q$. Expanding the MSE cost function:

$$J_k = \mathbb{E}[(s_k - \mathbf{v}_k^\dagger \mathbf{y}^q)(s_k - \mathbf{v}_k^\dagger \mathbf{y}^q)^\dagger] \tag{12}$$

$$= 1 - \mathbf{v}_k^\dagger \mathbb{E}[\mathbf{y}^q s_k^*] - \mathbb{E}[s_k(\mathbf{y}^q)^\dagger]\mathbf{v}_k + \mathbf{v}_k^\dagger \mathbf{R}_{\mathbf{y}^q}\mathbf{v}_k \tag{13}$$

Taking the gradient with respect to $\mathbf{v}_k^*$ and setting it to zero:

$$\nabla_{\mathbf{v}_k^*} J_k = -\mathbb{E}[\mathbf{y}^q s_k^*] + \mathbf{R}_{\mathbf{y}^q}\mathbf{v}_k = 0 \tag{14}$$

Thus, the optimal combining vector is $\mathbf{v}_k = \mathbf{R}_{\mathbf{y}^q}^{-1}\mathbb{E}[\mathbf{y}^q s_k^*]$. The cross-correlation vector is:

$$\mathbb{E}[\mathbf{y}^q s_k^*] = \mathbb{E}[(\alpha_B \mathbf{h}_k s_k + \dots)s_k^*] = \alpha_B \hat{\mathbf{h}}_k \tag{15}$$

The auto-correlation matrix $\mathbf{R}_{\mathbf{y}^q}$ is:

$$\mathbf{R}_{\mathbf{y}^q} = \alpha_B^2 \sum_{i=1}^{K} p_i \hat{\mathbf{h}}_i \hat{\mathbf{h}}_i^\dagger + \mathbf{C}_{\text{total\_noise}} \tag{16}$$

where $\mathbf{C}_{\text{total\_noise}}$ includes the thermal noise $\alpha_B^2 \sigma^2 \mathbf{I}$ and the quantization noise covariance $\mathbf{C}_\mathbf{q}$. Substituting these back yields the expression in Eq. (11).

REFERENCES

[1] M. Bashar, A. Akbari, K. Cumanan, H. Q. Ngo, A. G. Burr, P. Xiao, and M. Debbah, “Deep learning-aided finite-capacity fronthaul cell-free massive MIMO with zero forcing,” in *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, 2020, pp. 1–6, doi: 10.1109/ICC40277.2020.9149210.

[2] A. A. Alammari, M. Sharique, A. A. Moinuddin, and M. S. Ansari, “Local-partial signal combining schemes for cell-free large-scale MU-MIMO systems with limited fronthaul capacity and spatial correlation channels,” *Electronics*, vol. 11, no. 17, p. 2757, 2022, doi: 10.3390/electronics11172757.

[3] H. Masoumi and M. J. Emadi, “Performance analysis of cell-free massive MIMO system with limited fronthaul capacity and hardware impairments,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 2, pp. 1038–1053, 2020, doi: 10.1109/TWC.2019.2950316.

[4] Z. Li, F. Göttsch, S. Li, M. Chen, and G. Caire, “Joint fronthaul load balancing and computation resource allocation in cell-free user-centric massive MIMO networks,” *IEEE Transactions on Wireless Communications*, 2024, doi: 10.1109/TWC.2024.3409609.

[5] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, and M. Debbah, “Cell-free massive MIMO with limited backhaul,” in *2018 IEEE International Conference on Communications (ICC)*, 2018, pp. 1–7, doi: 10.1109/ICC.2018.8422865.

[6] A. Burr, M. Bashar, and D. Maryopi, "Cooperative access networks: Optimum fronthaul quantization in distributed massive mimo and cloud ran," in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, 2018, pp. 1–5, doi: 10.1109/VTCSpring.2018.8417560.

[7] I.-s. Kim, M. Bennis, and J. Choi, "Cell-free mmwave massive MIMO systems with low-capacity fronthaul links and low-resolution ADC/DACs," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 10, pp. 10 512–10 526, 2022, doi: 10.1109/TVT.2022.3184172.

[8] M. Kim, I. Kim, and J. Choi, "Meta-heuristic fronthaul bit allocation for cell-free massive MIMO systems," *IEEE Transactions on Wireless Communications*, 2024, doi: 10.1109/TWC.2024.3384548.

[9] I. Kanno, M. Ito, Y. Amano, Y. Kishi, T. Choi, W. Chen, and A. F. Molisch, "Adaptive bit allocation for SVD based hybrid processing of uplink cell-free massive MIMO under limited fronthaul capacity," in *2023 IEEE 97th Vehicular Technology Conference (VTC2023-Spring)*, 2023, pp. 1–5, doi: 10.1109/VTC2023-Spring57618.2023.10201009.

[10] I. Kanno, M. Ito, T. Ohseki, Y. Amano, and Y. Kishi, "Fronthaul load-reduced scalable cell-free massive MIMO by uplink hybrid signal processing," in *VTC2022-Spring: 2022 IEEE 95th Vehicular Technology Conference*, 2022, pp. 1–5, doi: 10.1109/VTC2022-Spring54318.2022.9860485.

[11] G. Femenias, F. Riera-Palou, and E. Björnson, "Another twist to the scalability in cell-free massive MIMO networks," *IEEE Transactions on Communications*, vol. 71, no. 11, pp. 6793–6804, 2023, doi: 10.1109/TCOMM.2023.3305531.

[12] S. Park, A. H. Gokceoglu, L. Wang, and O. Simeone, "Scalable multivariate fronthaul quantization for cell-free massive MIMO," *IEEE Transactions on Signal Processing*, vol. 73, pp. 1658–1673, 2025, doi: 10.1109/TSP.2025.3550469.

[13] M. Bashar, H. Q. Ngo, K. Cumanan, A. G. Burr, P. Xiao, E. Björnson, and E. G. Larsson, "Uplink spectral and energy efficiency of cell-free massive MIMO with optimal uniform quantization," *IEEE Transactions on Communications*, vol. 69, no. 1, pp. 223–245, 2021, doi: 10.1109/TCOMM.2020.3023253.

[14] Y. Khorsandmanesh, E. Björnson, and J. Jaldén, "Fronthaul quantization-aware MU-MIMO precoding for sum rate maximization," in *ICC 2023 - IEEE International Conference on Communications*, 2023, pp. 1332–1337, doi: 10.1109/ICC45041.2023.10279822.

[15] Y. Dong, Z. Yang, H. Wang, and N. Hao, "Efficient hybrid distributed detectors for cell-free massive MIMO systems," in *ICC 2024 - IEEE International Conference on Communications*, 2024, pp. 1–6, doi: 10.1109/ICC51166.2024.10622711.

[16] Z. Ren, A. Doufexi, and M. A. Beach, "Lattice quantization for centralized and scalable cell-free massive MIMO with realistic fronthaul," in *2022 IEEE International Mediterranean Conference on Communications and Networking (MeditCom)*, 2022, pp. 293–298, doi: 10.1109/MeditCom55741.2022.9928751.

[17] Y. Xiong, S. Sun, L. Qin, N. Wei, L. Liu, and Z. Zhang, "Performance analysis on cell-free massive MIMO with capacity-constrained fronthauls and variable-resolution ADCs," *IEEE Systems Journal*, vol. 16, no. 2, pp. 3296–3307, 2022, doi: 10.1109/JSYST.2021.3069753.

[18] G. Wang, Y. Gao, M. Guo, N. Sha, X. Wang, and N. Li, "Quantization analysis of cell-free massive MIMO uplink backhaul based on zero-forcing," in *2021 IEEE 21st International Conference on Communication Technology (ICCT)*, 2021, pp. 615–620, doi: 10.1109/ICCT52962.2021.9658000.

[19] M. Francis and K. V. S. Hari, "Selective user forwarded cell-free massive MIMO with quantized symbols," in *2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 8010–8014, doi: 10.1109/ICASSP48485.2024.10446251.

[20] J. Kim, H. Kim, K. Shin, S. Park, S. Seo, Y. Choi, Y. You, and H. Song, "User-centric cell-free massive MIMO with low-resolution ADCs for massive access," *Sensors*, vol. 24, no. 16, p. 5088, 2024, doi: 10.3390/s24165088.

[21] G. Interdonato, P. Frenger, and E. G. Larsson, "Ubiquitous cell-free massive MIMO communications," *EURASIP Journal on Wireless Communications and Networking*, vol. 2019, p. 197, 2019, doi: 10.1186/s13638-019-1507-0.

[22] K. Datta *et al.*, "Fd cell-free mmimo: Analysis and optimization," *International Conference on Communications*, 2021.

[23] E. V. Pothan and S. Kashyap, "Cell-free massive MIMO based underlay spectrum access under interference outage probability constraint over limited capacity fronthaul," *IEEE Transactions on Communications*, vol. 72, no. 9, pp. 5837–5852, 2024, doi: 10.1109/TCOMM.2024.3387853.

[24] H. He, H. Wang, X. Yu, J. Zhang, S. H. Song, and K. B. Letaief, "Distributed expectation propagation detection for cell-free massive mimo," in *2021 IEEE Global Communications Conference (GLOBECOM)*, 2021, pp. 01–06, doi: 10.1109/GLOBECOM46510.2021.9685334.

[25] S. Nassirpour, T.-V. Nguyen, H. Q. Ngo, L.-N. Tran, T. Ratnarajah, and D. H. N. Nguyen, "Variational bayesian channel estimation and data detection for cell-free massive mimo with low-resolution quantized fronthaul links," *IEEE Journal of Selected Topics in Signal Processing*, vol. 19, no. 6, pp. 1187–1202, 2025, doi: 10.1109/JSTSP.2025.3579644.

[26] M. Bashar, A. Akbari, K. Cumanan, H. Q. Ngo, A. G. Burr, P. Xiao, M. Debbah, and J. Kittler, "Exploiting deep learning in limited-fronthaul cell-free massive MIMO uplink," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 8, pp. 1678–1697, 2020, doi: 10.1109/JSAC.2020.3000812.

[27] D. Maryopi, M. Bashar, and A. Burr, "On the uplink throughput of zero forcing in cell-free massive MIMO with coarse quantization," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 7, pp. 7220–7224, 2019, doi: 10.1109/TVT.2019.2920070.

[28] O. Orhan, E. Erkip, and S. Rangan, "Low power analog-to-digital conversion in millimeter wave systems: Impact of resolution and bandwidth on performance," in *2015 IEEE Information Theory Workshop (ITW)*, 2015, pp. 1–5, doi: 10.1109/ITA.2015.7308988.

[29] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2325–2383, 1998, doi: 10.1109/18.720541.

[30] E. Björnson, M. Sanguinetti, J. Hoydis, and M. Debbah, "Optimal design of energy-efficient multi-user MIMO systems," *IEEE Transactions on Wireless Communications*, vol. 14, no. 8, pp. 4268–4281, 2015, doi: 10.1109/TWC.2015.2400437.

[31] T. Choi, I. Kanno, and A. F. Molisch, "Joint optimization of codebook and fronthaul compression for cell-free mmwave massive MIMO," in *2023 IEEE Globecom Workshops (GC Wkshps)*, 2023, pp. 1–6, doi: 10.1109/TVT.2022.3184172.