Ultra-Reliable and Low Latency Communication in mmWave-Enabled Massive MIMO Networks

Trung Kien Vu, Student Member, IEEE, Chen-Feng Liu, Student Member, IEEE, Mehdi Bennis, Senior Member, IEEE, Mérouane Debbah, Fellow, IEEE, Matti Latva-aho, Senior Member, IEEE, and Choong Seon Hong, Senior Member, IEEE

Abstract—Ultra-reliability and low latency are two key components in 5G networks. In this letter, we investigate the problem of ultra-reliable and low-latency communication in millimeter wave-enabled massive multiple-input multiple-output networks. The problem is cast as a network utility maximization subject to probabilistic latency and reliability constraints. To solve this problem, we resort to the Lyapunov technique, whereby a utility-delay control approach is proposed, which adapts to channel variations and queue dynamics. Numerical results demonstrate that our proposed approach ensures reliable communication with a guaranteed probability of 99.99%, and reduces latency by 28.41% and 77.11% as compared to baselines with and without probabilistic latency constraints, respectively.

Index Terms—5G, massive MIMO, mmWave, ultra-reliable low latency communications (URLLC).

I. INTRODUCTION

CURRENtLY, millimeter wave (mmWave) and massive multiple-input multiple-output (MIMO) techniques are investigated to provide reliable communication with an over-the-air latency of few milliseconds and extreme throughput [1]. While massive MIMO with large degrees of freedom provides high energy and spectral efficiency [2], mmWave frequency bands provide large bandwidth [3]. In addition, due to the short wavelength of mmWaves, large antenna array can be packed into highly directional beamforming, which makes massive MIMO practically feasible [4]. Thus far, most of existing works on mmWave-enabled massive MIMO systems focus mainly on providing capacity improvement [4], while latency and reliability are not addressed. Although latency and reliability are applicable to many scenarios (e.g., mission-critical applications), in this work, we are interested in the integration of mmWave communication and massive MIMO techniques, which holds the promise of providing great enhancements of the overall system performance [1], [2], [4]. Specifically, this letter is concerned with addressing the fundamental question in mmWave-enabled massive MIMO systems: "how to simultaneously provide order of magnitude capacity improvements and latency reduction?" By invoking the Lyapunov optimization framework, an utility-optimal solution is obtained to maximize network throughput subject to queuing stability [5]. This solution establishes a utility-delay tradeoff, which achieves utility-optimality at the price of large queuing delays [5]. To cope with this shortcoming, in this letter the Lyapunov framework is extended to incorporate probabilistic latency and reliability constraints, which takes into account queue length, arrival rate, and channel variations with a guaranteed probability. By applying the drift-plus-penalty technique, the problem is decoupled into a dynamic latency control and rate allocation. Here, the latency control problem is a difference of convex (DC) programming problem, which is solved efficiently by the convex-concave procedure (CCP) [6].

II. SYSTEM MODEL

Consider the downlink (DL) transmission of a single cell massive MIMO system1 consisting of one macro base station (MBS) equipped with $N$ antennas, and a set, $M = \{1, \ldots, M\}$, of single-antenna user equipments (UEs). We assume that $N \geq M$ and $N \gg 1$. Further, the co-channel time-division duplexing (TDD) is considered in which the MBS estimates channels via the uplink phase. We denote the propagation channel between the MBS and the $m$th UE as $h_m = \sqrt{\Theta_m} \hat{h}_m$, where $\Theta_m \in \mathbb{C}^{N \times N}$ depicts the antenna spatial correlation, and the elements of $\hat{h}_m \in \mathbb{C}^{N \times 1}$ are independent and identically distributed (i.i.d.) with zero mean and variance $1/N$. In addition, channels experience flat and block fading, and imperfect channel state information (CSI) is assumed. As per [9], the estimated channel can be modeled as $\hat{h}_m = \sqrt{1 - \tau_m^2} h_m + \tau_m \sqrt{\Theta_m} z_m, \forall m \in M$. Here, $z_m \in \mathbb{C}^{N \times 1}$ denotes the estimated noise vector which has i.i.d. elements with zero mean and variance $1/N$, and $\tau_m \in [0, 1]$ reflects the estimation accuracy; in case of perfect CSI, $\tau_m = 0$. Given the estimated channel matrix $\hat{H} = [\hat{h}_1, \ldots, \hat{h}_M] \in \mathbb{C}^{N \times M}$, the MBS utilizes the regularized zero-forcing (RZF) precoder with the preceding

---

1Our model can be extended to multi-cell massive MIMO systems in which the problem of inter-cell interference can be addressed by designing a hierarchical precoder at the MBS [7], to mitigate both intra-cell and inter-cell interference, or by applying an interference coordination approach [8].
matrix, \( V = [v_1, \ldots, v_M] \in \mathbb{C}^{N \times M} \), which is given by \( V = (\tilde{H}^H \tilde{H} + N\alpha I_N)^{-1} \tilde{H}^H \). Note that the regularization parameter \( \alpha > 0 \) is scaled by \( N \) to ensure the matrix \( \tilde{H}^H \tilde{H} + N\alpha I_N \) is well-conditioned as \( N \to \infty \) [7]. Denoting all allocated powers in the diagonal matrix \( P = \text{diag}(p_1, \ldots, p_M) \), we have the constraint \( \text{Tr}(P V^H V) \leq P \), with \( P \) the maximum transmit power of the MBS. With the aid of the results in [9, Th. 1], the transmit power constraint is

\[
\frac{1}{N} \sum_{m=1}^M \frac{p_m}{\lambda_m} \leq P, \quad \text{and } p_m \geq 0, \quad \forall m \in \mathcal{M},
\]

where the parameter \( \Omega_m \) is the solution to \( \Omega_m = \frac{1}{N} \text{Tr}(\Theta_m(\frac{1}{N} \sum_{m=1}^M \Theta_m + I_N)^{-1}) \). By designing the preceding matrix \( V \) and transmit power \( P \), the ergodic DL rate of UE \( m \in \mathcal{M} \) is \( r_m(P) = \mathbb{E}[\log_2(1 + (p_m \| v_m \|^2)/(1 + \sum_{m=1}^M p_m \| v_m \|^2))] \). Here, we invoke results from random matrix theory in order to get the deterministic equivalence for the ergodic DL rate [9]. In particular, as \( N \geq M \) and \( N \gg 1 \), for a small fixed \( \alpha \), the ergodic DL rate almost surely converges to \( r_m(P) \overset{a.s.}{\to} \log_2(1 + p_m(1 - \tau_m^2)) \), \( \forall m \in \mathcal{M} \), where \( \overset{a.s.}{\to} \) denotes almost sure convergence [7, 9, Th. 2]. Moreover, we assume that the MBS has queue buffers to store UE data [5]. The queue length for UE \( m \) at time slot \( t \) is denoted by \( Q_m(t) \) which evolves as follows:

\[
Q_m(t+1) = (Q_m(t) - r_m(t))^+ + a_m(t), \quad \forall m \in \mathcal{M},
\]

where \( x^+ = \max(x, 0) \), and \( a_m(t) \) is the data arrival rate of UE \( m \). Further, we assume that \( a_m(t) \) is i.i.d. over time slots with mean arrival rate \( \lambda_m \) and upper bounded by \( \rho_m \) [5].

### III. PROBLEM FORMULATION

According to Little’s law [10], the average delay is proportional to \( \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^T \mathbb{E}[Q_m(t)]/\lambda_m \). We use \( \mathbb{E}[Q_m(t)]/\lambda_m \) as a delay measure and enforce an upper bound \( d_m \). Note that the delay bound violation is related to reliability. Thus, taking into account the latency and reliability requirements, we characterize the delay bound violation with a tolerable probability. Specifically, we impose a probabilistic constraint on the queue size length for \( EM \in \mathcal{M} \) as follows:

\[
\Pr\left[ \frac{Q_m(t)}{\lambda_m} \geq d_m \right] \leq \epsilon_m, \quad \forall t.
\]

In (3), \( d_m \) reflects the UE delay requirement. Here, \( \epsilon_m \ll 1 \) is the target probability for reliable communication.

To reduce latency, the intuitive way is to send as many data as possible. However, this might over-allocate resources to UEs, i.e., \( r_m(t) \gg Q_m(t) \). Hence, we enforce a maximum rate constraint \( r_m^\text{max} \) for each UE \( m \). Moreover, we enforce the MBS to guarantee for all UEs a certain level of QoS, i.e., the minimum rate requirement \( r_m^\text{min} \), \( \forall m \in \mathcal{M} \).

We define the network utility as \( \sum_{m=1}^M \omega_m f(\bar{r}_m) \) where the time average expected rate \( \bar{r}_m = \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^T \mathbb{E}[r_m(t)] \) and the non-negative weight \( \omega_m \) for each UE \( m \). Additionally, we assume that \( f(\cdot) \) is a strictly concave, increasing, and twice continuously-differentiable function. Taking into account these constraints presented above yields the following network utility maximization:

\[
\text{OP : } \max_{P(t)} \sum_{m=1}^M \omega_m f(\bar{r}_m)
\]

subject to \( r_m^\text{min} \leq r_m(t) \leq r_m^\text{max}, \quad \forall m \in \mathcal{M}, \forall t, \quad (1) \) and (3).

Our main problem involves a probabilistic constraint (3), which cannot be addressed tractably. To overcome this challenge, we apply Markov’s inequality [11] to linearize (3) such that \( \Pr\left[ Q_m(t) \geq d_m \right] \leq \frac{\mathbb{E}[Q_m(t)]}{d_m} \). Then, (3) is satisfied if

\[
\mathbb{E}[Q_m(t)] \leq \lambda_m d_m \epsilon_m, \quad \forall m \in \mathcal{M}, \forall t.
\]

Thereafter, we consider (5) to represent the latency and reliability constraint. Assuming that \( \{a_m(t)|\forall t \geq 1\} \) is a Poisson arrival process [11], we have that \( \mathbb{E}[Q_m(t)] = t \lambda_m - \sum_{t=1}^{t-1} f_m(t) \) which is plugged into (5). Finally, we obtain

\[
r_m(t) \geq t \lambda_m - \lambda_m d_m \epsilon_m - \sum_{t=1}^{t-1} f_m(t), \quad \forall m \in \mathcal{M}, \forall t,
\]

which represents the minimum rate requirement in slot \( t \) for \( UE \) \( m \) for URLLC. Here, we transform the probabilistic latency and reliability constraint (3) into one linear constraint (6) of instantaneous rate requirements, which helps to analyse and optimize the URLLC problem. Combining (4b) and (6), we rewrite \( \text{OP} \) as follows:

\[
\text{OP : } \max_{P(t)} \sum_{m=1}^M \omega_m f(\bar{r}_m)
\]

subject to \( r_m^0(t) \leq r_m(t) \leq r_m^\text{max}, \quad \forall m \in \mathcal{M}, \forall t, \quad (1) \) and (7a).

with \( r_m^0(t) = \max\{r_m^\text{min}, t \lambda_m - \lambda_m d_m \epsilon_m - \sum_{t=1}^{t-1} f_m(t)\} \).

### IV. LYAPUNOV OPTIMIZATION FRAMEWORK

To tackle (7), we resort to Lyapunov framework [5]. Firstly, for each DL rate \( r_m(t) \), we introduce the auxiliary variable \( \phi_m(t) = (\varphi_m(t)|\forall m \in \mathcal{M}) \) that satisfies

\[
\phi_m(t) = \lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^T \mathbb{E}[\varphi_m(t)] \leq \bar{r}_m, \quad \forall m \in \mathcal{M}, \quad (8)
\]

with \( \varphi_m(t) = \max\{r_m^\text{min}, t \lambda_m - \lambda_m d_m \epsilon_m - \sum_{t=1}^{t-1} f_m(t)\} \).

Incorporating the auxiliary variables, (7) is equivalent to

\[
\text{LP : } \max_{P(t), \varphi(t)} \lim_{T \to \infty} \frac{1}{T} \sum_{t=1}^T \sum_{m=1}^M \omega_m \mathbb{E}[f(\varphi_m(t))] \]

subject to (1), (8), and (9).

In order to ensure the inequality constraint (8), a virtual queue vector \( Y(t) = (Y_m(t) | \forall m \in \mathcal{M}) \) is introduced, where each element evolves according to

\[
Y_m(t+1) = [Y_m(t) + \varphi_m(t) - r_m(t)]^+, \quad \forall m \in \mathcal{M}.
\]
Subsequently, we express the conditional Lyapunov drift-plus-penalty for each time slot \( t \) as:

\[
\mathbb{E} \left[ \sum_{m=1}^{M} \left( \frac{1}{2} Y_m(t+1)^2 - \frac{1}{2} Y_m(t)^2 - v_m(t) \omega_m f(\varphi_m(t)) \right) \right] Y(t) .
\]

(11)

In (11), \( v_m(t) \) is the control parameter which affects the utility-queue length tradeoff. This control parameter is conventionally chosen to be static and identical for all UEs [5]. However, this setting does not hold for system dynamics (e.g., instantaneous data arrivals) and the diverse system configuration (i.e., different delay and QoS requirements). Thus, we dynamically design these control parameters. From the analysis in the Lyapunov optimization framework [5], we can find \( Y_m(t) \leq v_m(t) \omega_m \pi_m + a_m^{\text{max}} \) with \( \pi_m \) being the largest first-order derivative of \( f(x) \). Letting \( \omega_m = 1, \forall m \in M \), we have the lower bound \( \pi_m v_m \geq v_0(t), \forall m \in M \), for selecting the control parameters, where \( v_0(t) = \max\{Y_m(t) - a_m^{\text{max}}, 1\} \). Subsequently, following the straightforward calculations of the Lyapunov drift-plus-penalty technique we obtain

\[
(11) \leq \mathbb{E} \left[ \sum_{m=1}^{M} \left( Y_m(t) \varphi_m(t) - v_m(t) \omega_m f(\varphi_m(t)) \right) \right] (12a)
- \sum_{m=1}^{M} Y_m(t) r_m(P(t)) + C \left( Y(t) \right) .
\]

(12b)

Due to space limitation, we omit the details of the constant value \( C \) which does not influence the system performance [5].

We note that the solution to LP is acquired by minimizing the right-hand side (RHS) of (12a) and (12b) in every slot \( t \). Further, (12a) is related to the reliability and QoS requirements while (12b) reflects optimal power allocation to UEs.

A. Auxiliary Variable and Control Parameter Selection

Considering the logarithmic fairness utility function, i.e., \( f(x) = \log(x) \), minimizing the RHS of (12a) for each \( m \in M \) is formulated as:

\[
\begin{align*}
\min_{\varphi_m(t), v_m(t)} & \quad Y_m(t) \varphi_m(t) - v_m(t) \log(\varphi_m(t)) \\
\text{subject to} & \quad \pi_m v_m \geq v_0(t), \\
& \quad r_m \leq \varphi_m(t) \leq r_m^{\text{max}}. \quad (13a)
\end{align*}
\]

Before proceeding with (13), we rewrite \( v_m(t) \log(\varphi_m(t)) \) in (13a), for any \( \varphi_m(t) > 0 \) and \( v_m(t) > 0 \), as

\[
\frac{v_m(t) \log(\varphi_m(t)) - v_m(t) \log(\varphi_m(t))}{h_0(\varphi_m, v_m)} \cdot g_0(v_m)
\]

in which both \( h_0(\varphi_m, v_m) \) (i.e., relative entropy function) and \( g_0(v_m) \) (i.e., negative entropy function) are convex functions.

Since (13a) is the difference of convex functions while constraints (13b) and (13c) are affine functions, problem (13) belongs to DC programming [12], which can be efficiently and iteratively addressed by the CCP [6]. The CCP algorithm to obtain the solution to problem (13) is detailed in Algorithm 1, which provably converges to the local optima of DC programming [6] (please refer to [6] for the formal proof).

Algorithm 1 CCP Algorithm for Solving Sub-Problem (13)

\[
\text{while } m \in M \text{ do}
\]

\[
\text{Initialize } i = 0 \text{ and a feasible point } v_m^{(i)} \text{ in (13b).}
\]

\[
\text{repeat}
\]

\[
\text{Convexify } \tilde{g}_0(v_m, v_m^{(i)}) = g_0(v_m^{(i)}) + \nabla g_0(v_m^{(i)}).
\]

\[
\text{Solve:}
\]

\[
\begin{align*}
\min_{\varphi_m, v_m} & \quad h_0(\varphi_m, v_m) - \tilde{g}_0(v_m, v_m^{(i)}) + Y_m \varphi_m \\
\text{subject to} & \quad (13b) \text{ and (13c)}, \quad \text{Find the optimal } \varphi_m^{(i+1)} \text{ and } v_m^{(i+1)} . \quad \text{Update } v_m^{(i+1)} \text{ and } i := i + 1. \\
\text{until Convergence}
\end{align*}
\]

B. Power Allocation

The optimal transmit power in (12b) is computed by

\[
\min_{P(t)} \quad - \sum_{m=1}^{M} Y_m(t) r_m(P(t)) \leq 0 \quad \text{subject to } (1).
\]

Here, the objective function is strictly convex for \( p_m(t) \geq 0, \forall m \in M \), and the constraints are compact. Therefore, the optimal solution of \( P^*(t) \) exists.

After obtaining the optimal auxiliary variable and transmit power, we update the queues \( Q_m(t+1) \) and \( Y_m(t+1) \) as per (2) and (10), respectively.

V. NUMERICAL RESULTS

We consider a single-cell massive MIMO system in which the MBS, with \( N = 32 \) antennas and \( P = 38 \) dBm, is located at the center of the \( 0.5 \times 0.5 \text{ km}^2 \) square area. UEs (from 8 to 60 UEs per \( \text{km}^2 \)) are randomly deployed within the MBS’s coverage with a minimum MBS-UE distance of 35 m. Data arrivals follow a Poisson distribution with different means, and the rate requirements are specified as \( r_m^{\text{max}} = 1.2 \lambda_m, r_m^{\text{min}} = 0.8 \lambda_m, \forall m \in M \). The system bandwidth is 1 GHz. The path loss is modeled as a distance-based path loss with the line-of-sight (LOS) model for urban environments at 28 GHz [13].

\[
\begin{align*}
\lambda^\text{th} & \quad \text{and } \epsilon \text{ are set to } 10 \text{ ms and } 5\%, \text{ respectively. The numerical results are obtained via Monte-Carlo simulations over 10,000 channel realizations. Furthermore, we compare our proposed scheme with the following baselines:}
\end{align*}
\]

1) Baseline 1 refers to the Lyapunov framework in which the probabilistic latency constraint (3) is considered.

2) Baseline 2 is a variant of Baseline 1 without the probabilistic latency constraint (3).

A. Impact of Arrival Rate

In Fig. 1, we report the average latency versus the mean arrival rates \( \lambda = \mathbb{E}[a(t)] \) for \( M = 16 \). At low \( \lambda \), all schemes do not violate latency constraints, and our proposed algorithm outperforms other baselines with a small gap. At higher \( \lambda \), the average delay of baseline 2 increases dramatically as \( \lambda > 1.8 \text{ Gbps} \), since baseline 2 does not incorporate the delay constraint, whereas our proposed scheme reduces latency by 28.4% and 77.1% as compared to baselines 1 and 2, respectively, when \( \lambda = 2.4 \text{ Gbps} \). When \( \lambda > 2.4 \text{ Gbps} \), the average delay of all schemes increases, violating the delay requirement...
and average latency of our proposed approach with the baseline 2 and baseline 3 cases. The delay time decreases with an increase in the number of users. This highlights the trade-off between latency and network density. The proposed approach achieves better throughput and higher latency reduction than baselines 1 and 2, while the WSRM case has the worst delay performance as expected. Moreover, our proposed approach reaches Gbps capacity, which represents the capacity improvement brought by the combination of mmWave and massive MIMO techniques.

VI. CONCLUSION

In this letter, we have investigated the problem of mmWave-enabled massive MIMO networks from a latency and reliability standpoint. Specifically, the problem is modeled as a NUM problem subject to the probabilistic latency/reliability constraint and QoS/rate requirement. Numerical results show that our proposed approach reduces the latency by 28.41% and 77.11% as compared to current baselines.

REFERENCES


