

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*, M erouane 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, $\mathcal{M} = \{1, \dots, M\}$, 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 $\mathbf{h}_m = \sqrt{N}\mathbf{\Theta}_m^{1/2}\tilde{\mathbf{h}}_m$, where $\mathbf{\Theta}_m \in \mathbb{C}^{N \times N}$ depicts the antenna spatial correlation, and the elements of $\tilde{\mathbf{h}}_m \in \mathbb{C}^{N \times 1}$ 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 $\hat{\mathbf{h}}_m = \sqrt{1 - \tau_m^2}\mathbf{h}_m + \tau_m\sqrt{N}\mathbf{\Theta}_m^{1/2}\mathbf{z}_m, \forall m \in \mathcal{M}$. Here, $\mathbf{z}_m \in \mathbb{C}^{N \times 1}$ 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 $\hat{\mathbf{H}} = [\hat{\mathbf{h}}_1, \dots, \hat{\mathbf{h}}_M] \in \mathbb{C}^{N \times M}$, the MBS utilizes the regularized zero-forcing (RZF) precoder with the precoding

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].

Manuscript received May 7, 2017; accepted May 12, 2017. Date of publication May 17, 2017; date of current version September 8, 2017. This work was supported in part by the Finnish Funding Agency for Technology and Innovation (Tekes), Nokia, Huawei, Anite, in part by the Academy of Finland CARMA project, in part by the Academy of Finland funding through the Grant 284704, and in part by the ERC Starting Grant 305123 MORE (Advanced Mathematical Tools for Complex Network Engineering). The associate editor coordinating the review of this letter and approving it for publication was L. Wang. (*Corresponding author: Trung Kien Vu.*)

T. K. Vu, C.-F. Liu, and M. Latva-aho are with the Centre for Wireless Communications, University of Oulu, 90014 Oulu, Finland (e-mail: trungkien.vu@oulu.fi; chen-feng.liu@oulu.fi; matti.latva-aho@oulu.fi).

M. Bennis is with the Centre for Wireless Communications, University of Oulu, 90014 Oulu, Finland, and also with the Department of Computer Engineering, Kyung Hee University, Yongin 446-701, South Korea (e-mail: mehdi.bennis@oulu.fi).

M. Debbah is with the Large Networks and System Group, CentraleSup elec, Universit  Paris-Saclay, 91192 Gif-sur-Yvette, France, and also with the Mathematical and Algorithmic Sciences Laboratory, Huawei France Research and Development, 92100 Paris, France (e-mail: merouane.debbah@huawei.com).

C. S. Hong is with the Department of Computer Engineering, Kyung Hee University, Yongin 446-701, South Korea (e-mail: cshong@khu.ac.kr).

Digital Object Identifier 10.1109/LCOMM.2017.2705148

matrix, $\mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_M] \in \mathbb{C}^{N \times M}$, which is given by $\mathbf{V} = (\hat{\mathbf{H}}^\dagger \hat{\mathbf{H}} + N\alpha \mathbf{I}_N)^{-1} \hat{\mathbf{H}}^\dagger$. Note that the regularization parameter $\alpha > 0$ is scaled by N to ensure the matrix $\hat{\mathbf{H}}^\dagger \hat{\mathbf{H}} + N\alpha \mathbf{I}_N$ is well-conditioned as $N \rightarrow \infty$ [7]. Denoting all allocated powers in the diagonal matrix $\mathbf{P} = \text{diag}(p_1, \dots, p_M)$, we have the constraint $\text{Tr}(\mathbf{P}\mathbf{V}^\dagger\mathbf{V}) \leq P$, with P the maximum transmit power of the MBS. With the aid of the results in [9, Th. 1], the transmit power constraint is

$$\frac{1}{N} \sum_{m=1}^M \frac{p_m}{\Omega_m} \leq P, \quad \text{and } p_m \geq 0, \quad \forall m \in \mathcal{M}, \quad (1)$$

where the parameter Ω_m is the solution to $\Omega_m = \frac{1}{N} \text{Tr}(\Theta_m (\frac{1}{N} \sum_{m=1}^M \frac{\Theta_m}{\alpha + \Omega_m} + \mathbf{I}_N)^{-1})$. By designing the precoding matrix \mathbf{V} and transmit power \mathbf{P} , the ergodic DL rate of UE $m \in \mathcal{M}$ is $r_m(\mathbf{P}) = \mathbb{E}[\log_2(1 + (p_m |\mathbf{h}_m^\dagger \mathbf{v}_m|^2) / (1 + \sum_{k=1, k \neq m}^M p_k |\mathbf{h}_m^\dagger \mathbf{v}_k|^2))]$. 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 α , the ergodic DL rate *almost surely* converges to $r_m(\mathbf{P}) \xrightarrow{a.s.} \log_2(1 + p_m(1 - \tau_m^2))$, $\forall m \in \mathcal{M}$, where $\xrightarrow{a.s.}$ 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:

$$Q_m(t+1) = [Q_m(t) - r_m(t)]^+ + a_m(t), \quad \forall m \in \mathcal{M}, \quad (2)$$

where $[x]^+ \triangleq \max\{x, 0\}$, 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 λ_m and upper bounded by a_m^{\max} [5].

III. PROBLEM FORMULATION

According to Little's law [10], the average delay is proportional to $\lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \mathbb{E}[Q_m(t)] / \lambda_m$. We use $Q_m(t) / \lambda_m$ as a delay measure and enforce an allowable upper bound d_m^{th} . 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 UE $m \in \mathcal{M}$ as follows:

$$\Pr\left\{\frac{Q_m(t)}{\lambda_m} \geq d_m^{\text{th}}\right\} \leq \epsilon_m, \quad \forall t. \quad (3)$$

In (3), d_m^{th} 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^{\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^{\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 $\bar{r}_m = \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \mathbb{E}[r_m(t)]$ and the non-negative weight ω_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:

$$\text{OP} : \max_{\mathbf{P}(t)} \sum_{m=1}^M \omega_m f(\bar{r}_m) \quad (4a)$$

$$\text{subject to } r_m^{\min} \leq r_m(t) \leq r_m^{\max}, \quad \forall m \in \mathcal{M}, \forall t, \\ (1) \text{ and } (3). \quad (4b)$$

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 $\Pr\left\{\frac{Q_m(t)}{\lambda_m} \geq d_m^{\text{th}}\right\} \leq \frac{\mathbb{E}[Q_m(t)]}{\lambda_m d_m^{\text{th}}}$. Then, (3) is satisfied if

$$\mathbb{E}[Q_m(t)] \leq \lambda_m d_m^{\text{th}} \epsilon_m, \quad \forall m \in \mathcal{M}, \forall t. \quad (5)$$

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 $\mathbb{E}[Q_m(t)] = t\lambda_m - \sum_{\tau=1}^t r_m(\tau)$ which is plugged into (5). Finally, we obtain

$$r_m(t) \geq t\lambda_m - \lambda_m d_m^{\text{th}} \epsilon_m - \sum_{\tau=1}^{t-1} r_m(\tau), \quad \forall m \in \mathcal{M}, \forall t, \quad (6)$$

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 **OP** as follows:

$$\max_{\mathbf{P}(t)} \sum_{m=1}^M \omega_m f(\bar{r}_m) \quad (7a)$$

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

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

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 vector $\boldsymbol{\varphi}(t) = (\varphi_m(t) | \forall m \in \mathcal{M})$ that satisfies

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

$$\varphi_m^0(t) \leq \varphi_m(t) \leq r_m^{\max}, \quad \forall m \in \mathcal{M}, \forall t, \quad (9)$$

with $\varphi_m^0(t) = \max\{r_m^{\min}, t\lambda_m - \lambda_m d_m^{\text{th}} \epsilon_m - \sum_{\tau=1}^{t-1} \varphi_m(\tau)\}$. Incorporating the auxiliary variables, (7) is equivalent to

$$\text{LP} : \max_{\mathbf{P}(t), \boldsymbol{\varphi}(t)} \lim_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T \sum_{m=1}^M \omega_m \mathbb{E}[f(\varphi_m(t))] \\ \text{subject to } (1), (8), \text{ and } (9).$$

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

$$Y_m(t+1) = [Y_m(t) + \varphi_m(t) - r_m(t)]^+, \quad \forall m \in \mathcal{M}. \quad (10)$$

Subsequently, we express the conditional Lyapunov drift-plus-penalty for each time slot t as:

$$\mathbb{E} \left[\sum_{m=1}^M \left[\frac{1}{2} Y_m(t+1)^2 - \frac{1}{2} Y_m(t)^2 - v_m(t) \omega_m f(\varphi_m(t)) \right] | \mathbf{Y}(t) \right]. \quad (11)$$

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 $Y_m(t) \leq v_m(t) \omega_m \pi_m + a_m^{\max}$ with π_m being the largest first-order derivative of $f(x)$. Letting $\omega_m = 1, \forall m \in \mathcal{M}$, we have the lower bound $\pi_m v_m(t) \geq v_m^0(t), \forall m \in \mathcal{M}$, for selecting the control parameters, where $v_m^0(t) = \max\{Y_m(t) - a_m^{\max}, 1\}$. Subsequently, following the straightforward calculations of the Lyapunov drift-plus-penalty technique we obtain

$$(11) \leq \mathbb{E} \left[\sum_{m=1}^M (Y_m(t) \varphi_m(t) - v_m(t) \omega_m f(\varphi_m(t))) \right] \quad (12a)$$

$$- \sum_{m=1}^M Y_m(t) r_m(\mathbf{P}(t)) + C | \mathbf{Y}(t) \right]. \quad (12b)$$

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 $m \in \mathcal{M}$ is formulated as:

$$\min_{\varphi_m(t), v_m(t)} Y_m(t) \varphi_m(t) - v_m(t) \log(\varphi_m(t)) \quad (13a)$$

$$\text{subject to } \pi_m v_m(t) \geq v_m^0(t), \quad (13b)$$

$$r_m^0(t) \leq \varphi_m(t) \leq r_m^{\max}. \quad (13c)$$

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

$$\underbrace{v_m(t) \log\left(\frac{v_m(t)}{\varphi_m(t)}\right)}_{h_0(\varphi_m, v_m)} - \underbrace{v_m(t) \log(v_m(t))}_{g_0(v_m)},$$

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)

while $m \in \mathcal{M}$ **do**

Initialize $i = 0$ and a feasible point $v_m^{(i)}$ in (13b).

repeat

Convexify $\hat{g}_0(v_m, v_m^{(i)}) = g_0(v_m^{(i)}) + \nabla g_0(v_m - v_m^{(i)})$.

Solve:

$$\min_{\varphi_m, v_m} h_0(\varphi_m, v_m) - \hat{g}_0(v_m, v_m^{(i)}) + Y_m \varphi_m$$

subject to (13b) and (13c),

Find the optimal $\varphi_m^{(i)*}$ and $v_m^{(i)*}$.

Update $v_m^{(i+1)} := v_m^{(i)*}$ and $i := i + 1$.

until Convergence

end while

B. Power Allocation

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

$$\min_{\mathbf{P}(t)} - \sum_{m=1}^M Y_m(t) r_m(\mathbf{P}(t))$$

subject to (1).

Here, the objective function is strictly convex for $p_m(t) \geq 0$, $\forall m \in \mathcal{M}$, and the constraints are compact. Therefore, the optimal solution of $\mathbf{P}^*(t)$ 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 0.5×0.5 km² square area. UEs (from 8 to 60 UEs per km²) 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 $r_m^{\max} = 1.2\lambda_m$, $r_m^{\min} = 0.8\lambda_m, \forall m \in \mathcal{M}$. 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]. d^{th} and ϵ are set to 10 ms and 5%, 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:

- 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 λ , all schemes do not violate latency constraints, and our proposed algorithm outperforms other *baselines* with a small gap. At higher λ , the average delay of *baseline 2* increases dramatically as $\lambda > 1.8$ Gbps, since *baseline 2* does not incorporate the delay constraint, whereas our proposed scheme reduces latency by 28.41% and 77.11% as compared to *baselines 1* and *2*, respectively, when $\lambda = 2.4$ Gbps. When $\lambda > 2.4$ Gbps, the average delay of all schemes increases, violating the delay requirement

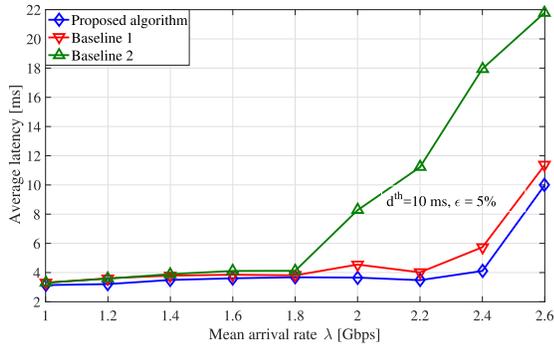


Fig. 1. Average latency versus mean arrival rates, $M = 16$ per km^2 .

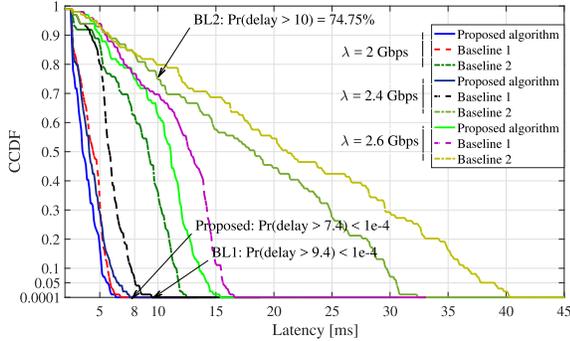


Fig. 2. Tail distribution (CCDF) of latency.

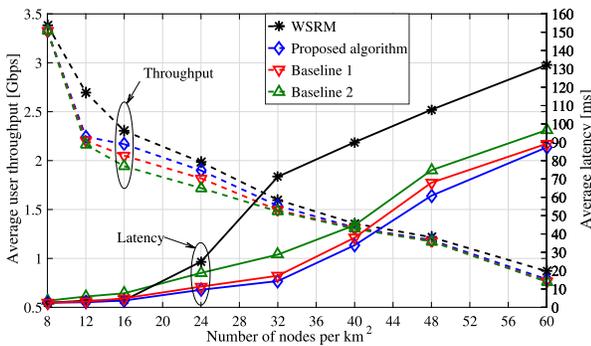


Fig. 3. Average latency and avgUT versus number of users per km^2 .

of 10ms. It can be observed that under a limited maximum transmit power, at very high traffic demand, the latency requirement could not be guaranteed. This highlights the trade-off between the arrival rate and latency. In Fig. 2, we report the tail distribution (complementary cumulative distribution function (CCDF)) of latency to showcase how often the system achieves a delay greater than target delay levels. In particular, at $\lambda = 2.4$ Gbps, by imposing the probabilistic latency constraint (3), our proposed approach and *baseline 1* ensure reliable communication with better guaranteed probabilities, i.e., $\Pr(\text{delay} > 7.5\text{ms}) < 10^{-4}$ and $\Pr(\text{delay} > 9.4\text{ms}) < 10^{-4}$, respectively. In contrast, *baseline 2* violates the latency constraint with a high probability, where $\Pr(\text{delay} > 10\text{ms}) = 74.75\%$.

B. Impact of User Density

In Fig. 3, we compare the average user throughput (avgUT) and average latency of our proposed approach with the two

baselines under the impact of user density, when $\lambda = 2$ Gbps. Additionally, we consider the weighted sum rate maximization (WSRM) case. The WSRM case is to find the system throughput limit but suffers from higher latency. Since all users share the same resources, the average delay (“solid lines”) increases with the number of users M , whereas the avgUT (“dash lines”) decreases. Fig. 3 further shows that when $M > 24$, the delay of all schemes increases dramatically and is far-above the latency requirement. Hence, only a limited number of users can be served to guarantee the delay requirement, above which, a tradeoff between latency and network density exists. Our 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

- [1] Nokia Siemens Networks, “2020: Beyond 4G: Radio evolution for the gigabit experience,” Espoo, Finland, White Paper, 2011.
- [2] E. G. Larsson *et al.*, “Massive MIMO for next generation wireless systems,” *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 186–195, Feb. 2014.
- [3] O. Semiari *et al.*, “Downlink cell association and load balancing for joint millimeter wave-microwave cellular networks,” in *Proc. IEEE Global Commun. Conf.*, Washington, DC, USA, Dec. 2016, pp. 1–6.
- [4] T. K. Vu *et al.*, “Joint in-band backhauling and interference mitigation in 5G heterogeneous networks,” in *Proc. 22nd Eur. Wireless Conf.*, Oulu, Finland, May 2016, pp. 1–6.
- [5] M. J. Neely, *Stochastic Network Optimization With Application to Communication and Queueing Systems*. San Mateo, CA, USA: Morgan & Claypool, 2010.
- [6] T. Lipp and S. Boyd, “Variations and extension of the convex-concave procedure,” *Optim. Eng.*, vol. 17, no. 2, pp. 263–287, 2016.
- [7] A. Liu and V. Lau, “Hierarchical interference mitigation for massive MIMO cellular networks,” *IEEE Trans. Signal Process.*, vol. 62, no. 18, pp. 4786–4797, Sep. 2014.
- [8] W. Feng *et al.*, “When mmWave communications meet network densification: A scalable interference coordination perspective,” *IEEE J. Sel. Areas Commun.*, to be published.
- [9] S. Wagner *et al.*, “Large system analysis of linear precoding in correlated MISO broadcast channels under limited feedback,” *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4509–4537, Jul. 2012.
- [10] J. D. C. Little and S. C. Graves, “Little’s law,” in *Building Intuition*. New York, NY, USA: Springer, 2008, pp. 81–100.
- [11] A. Mukherjee, “Queue-aware dynamic on/off switching of small cells in dense heterogeneous networks,” in *Proc. IEEE Global Commun. Conf. Workshops*, Atlanta, GA, USA, Dec. 2013, pp. 182–187.
- [12] L. T. H. An and P. D. Tao, “The DC (difference of convex functions) programming and DCA revisited with DC models of real world nonconvex optimization problems,” *Ann. Oper. Res.*, vol. 133, no. 1, pp. 23–46, Jan. 2005.
- [13] M. R. Akdeniz *et al.*, “Millimeter wave channel modeling and cellular capacity evaluation,” *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1164–1179, Jun. 2014.