IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 20, 2022, accepted February 25, 2022, date of publication March 2, 2022, date of current version March 30, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3156111

Latency-Aware Highly-Reliable mmWave Systems via Multi-Point Connectivity

DILEEP KUMAR, (Graduate Student Member, IEEE), SATYA KRISHNA JOSHI, (Member, IEEE), AND ANTTI TÖLLI, (Senior Member, IEEE)
Centre for Wireless Communications, University of Oulu, 90014 Oulu, Finland

Corresponding author: Dileep Kumar (dileep.kumar@oulu.fi)

This work was supported in part by the European Commission in the Framework of the H2020-EUJ-02-2018 Project (5G-Enhance) under Grant 815056, and in part by the Academy of Finland (6Genesis Flagship) under Grant 318927.

ABSTRACT The sensitivity of the millimeter-wave (mmWave) radio channel to blockages is a fundamental challenge in achieving low-latency and highly-reliable connectivity. In this paper, we explore the viability of using Coordinated Multi-Point (CoMP) transmission for delay bounded and reliable mmWave systems. We propose a blockage-aware framework for the sum-power minimization problem under the user-specific latency requirements in time dynamic mobile access networks. We use the Lyapunov optimization approach and provide a dynamic control algorithm, which transforms a time-average stochastic problem into a sequence of deterministic subproblems. A robust beamformer design is then proposed by exploiting the queue backlogs and channel information, that efficiently allocates the required resources, by proactively tuning the CoMP subsets from the available remote radio units (RRUs), according to the instantaneous needs of the users. Further, to adapt to the uncertainties of the mmWave channel, we consider a pessimistic estimate of the rates over link blockage combinations across the CoMP serving set. Moreover, after the relaxation of coupled and non-convex constraints via the Fractional Program (FP) techniques, a low-complexity closed-form iterative algorithm is provided by solving a system of Karush-Kuhn-Tucker (KKT) optimality conditions. The simulation results manifest that, in the presence of random link blockages, the proposed methods outperform the baseline scenarios and provide power-efficient, highly-reliable, and low-latency mmWave communication.

INDEX TERMS JT-CoMP, reliable communication, queue backlogs, sum-power minimization, Lyapunov framework, convex optimization, Karush-Kuhn-Tucker conditions.

I. INTRODUCTION

The millimeter-wave (mmWave) and sub-terahertz (sub-THz) communication are one of the key enabling technologies for 5th-generation (5G) and beyond cellular systems, which facilitates throughput-intensive and low-latency applications, such as Industrial Internet-of-Things (IIoT), factory automation, augmented reality, and autonomous driving [2]. However, the full exploitation of the large available bandwidth at higher frequencies, is mainly challenged by the sensitivity of directional radio links to the blockage, i.e., due to relatively higher penetration and path-losses. These lead to rapid degradation (i.e., strong dips) in the received signal strength, and thus result in intermittent connectivity [3], [4]. For example, a human blocker can obstruct the dominant links for hundred of milliseconds, and may lead to disconnecting the ongoing communication session, which severely impacts the network's reliability [3], [4]. Moreover, adapting to unpredictable blockage demands critical latency and signalling overhead, i.e., searching for an unblocked direction to re-establish the communication link [5]. Therefore, unless being addressed properly, the blockage appears as the main bottleneck that hinders the full exploitation of radio resources and in achieving highly-reliable low-latency connectivity.

To tackle the mmWave radio channel uncertainties, the use of macro-diversity via Coordinated Multi-Point (CoMP) has gained great interest. In particular, the Joint Transmission (JT)-CoMP connectivity, where each user equipment (UE) is coherently served by multiple spatially distributed remote radio units (RRUs) [6], [7]. Further, CoMP schemes

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

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

are also reconsidered in recent 3rd Generation Partnership Project (3GPP) releases [8], and it is envisioned that the use of multi-antenna spatial redundancies via geographically separated transceivers will be of high importance in the future mmWave and sub-THz based deployment scenarios. As a proof-of-concept, Qualcomm has recently implemented a CoMP testbed at mmWave frequencies [9], and empirically shown the coverage and capacity improvements via flexible deployments in IIoT and factory automation scenarios.

A. PRIOR WORK

The CoMP schemes are mainly employed to enhance the system throughput, generally for the cell-edge users due to adverse channel conditions (e.g., higher path-loss and interference from neighboring RRUs). The CoMP techniques, such as, JT, Coordinated Beamforming (CB) and Dynamic Point Selection (DPS) are standardized in 3GPP [8], and has been widely studied in past decade under the context of legacy 4G systems [10]–[12]. For example, in [12], it is shown that JT-CoMP increases the coverage by up to 24% for cell-edge users and 17% for general users compared to non-cooperative scenarios. Recent studies have also considered the CoMP schemes in the mmWave frequencies [13]–[16].

In [13], from extensive real-time measurements, the authors showed a significant coverage improvement by simultaneously serving a user with spatially distributed transmitters. The network coverage gain for the mmWave system with multi-point connectivity, in the presence of random blockages, was also confirmed in [14], [15] using stochastic geometry tools. The gains of macro-diversity for the achievable rate and outage probability were quantified in [16], and it was shown that CoMP connectivity, at the minimum of four spatially distributed links, can provide up to 76% capacity gains. However, CoMP schemes for the emerging wireless systems were still devised with the sole scope of enhancing the capacity and coverage [13]–[16], e.g., by efficiently utilizing the multi-antenna spatial redundancies via spatially separated RRUs. Thus, these techniques are not originally designed for the stringent latency and reliability requirements, which are inherent for industrial-grade critical applications. A step towards this direction is introduced in our earlier work [6], where we provide reliable CoMP transmission schemes in the presence of random blockages, by preemptively underestimating the achievable rates over the potential link blockage combinations. However, these algorithms are still designed for the time-invariant and static case, i.e., the resource allocation problem for a given instance is studied without taking into account network dynamics and stringent latency conditions due to data arrivals and evolving queue backlogs. Hence, these algorithms are not always applicable for, e.g., long-term time-dependent dynamics networks, and for retaining a robust and resilient mmWave connectivity while satisfying the user-specific latency requirements.

Thus, the limitations of retransmission events for, e.g., delay bounded critical applications [17], [18] and the difficulty of accurate estimation of random blocking events motivate us to develop latency-constrained highly-reliable transmission strategies for time dynamic networks. Specifically, we investigate on, how to use queue backlogs and channel information at the transmitter to efficiently allocate the required radio and cooperation resources, and to proactively exploit the multi-antenna spatial diversity according to the instantaneous needs of the users for dynamic mmWave access networks.

B. CONTRIBUTIONS

We develop a robust downlink transmission strategy, tailored for a JT-CoMP based dynamic networks, satisfying the user-specific latency requirements while retaining stable and resilient connectivity under the uncertainties of mmWave radio channel. Specifically, we consider an average sum-power minimization problem subject to maximum allowable queue length constraint per user in the presence of random link blockages. The long-term time-average stochastic problem is transformed into a sequence of deterministic and independent subproblems using the Lyapunov optimization framework [19]. The coupled and non-convex constraints are approximated with the sequence of convex subsets by using the Fractional Program (FP) techniques [20]. Further, the proposed FP based relaxations allow an efficient implementation of closed-form iterative beamformer design that enables tailored complexity and processing performance. The main contributions of this paper are summarized as follow:

- A robust transmit beamformer design is proposed by utilizing the multi-antenna spatial diversity and geographically separated transceivers in CoMP connectivity scenarios. The average sum-power is minimized while ensuring the latency requirements, where, for each user, a pessimistic estimate of the rates overall possible subset combinations of potentially blocked links is considered [6]. Thus, managing mutually coupled link blocked combinations is more challenging than conventional constrained optimization.
- To adapt with the uncertainties of the mmWave radio channel, a proactive and dynamic selection of the user-specific CoMP serving set from the available RRUs is proposed, e.g., by exploiting the queue backlogs and channel information. This preemptive rate estimate and dynamic selection of the serving subset is shown to greatly improve the reliability and average sum-power performance while ensuring the latency requirements.
- After the relaxation of coupled and non-convex constraints via the Fractional Program (FP) techniques, a low-complexity robust beamformer design framework is proposed by solving a system of closed-form Karush-Kuhn-Tucker (KKT) optimality conditions, and does not require any generic (or tailored) convex solvers. This leads to practical and computationally efficient implementation for, e.g., hardware-constrained devices with limited processing capabilities.

The paper is an extended version of our earlier published conference paper [1]. In this paper, we have included the following additional notable contributions that provide more complete coverage and analysis. We propose an algorithm for dynamic selection of user-specific CoMP serving subset, which ensures the latency requirements with the minimum sum-power, and efficiently adopts with the uncertainties of mmWave radio channel. Specifically, leveraging the queue backlogs and channel information at transmitter, we propose a latency-reliability aware framework to efficiently allocate the required radio and CoMP resources, and to proactively exploit spatial diversity according to the instantaneous needs of the users in time dynamic mobile network conditions.

We extend the FP quadratic transform techniques [20], i.e., to take into consideration JT-CoMP transmission and provide a novel grouping of a multitude of potentially coupled and non-convex SINR conditions, that raise from the link blockage subset combinations of CoMP serving set. Our proposed extension to FP facilitates a low-complexity iterative algorithm, where each step is efficiently computed in closed-form expressions, and thus, enables tailored complexity and convergence performance. All of the aforementioned results have been further improved and extended in this paper while accounting network dynamics and random link blockages. Finally, we provide comprehensive and detailed numerical examples to evaluate the performance advantage of the proposed solutions. Specifically, we quantify the underlying trade-offs in terms of average sum-power, achievable rates, and reliable downlink transmissions, under the uncertainties of mmWave radio channel and random link blockages.

C. ORGANIZATION AND NOTATIONS

The remainder of this paper is organized as follows. In Section II, we illustrate the system, blockage, and queuing model, as well as, provide the formulation of the problem. Section III provides a dynamic control algorithm. In Section IV, we describe the proposed beamformer designs. Section V provides an algorithm for dynamic serving set selection and theoretical analysis of outage. The validation of our proposed methods with the numerical results are presented in Section VI, and conclusions are given in Section VII.

Notations: In the following, we denote vectors and matrices with boldface lowercase and uppercase letters, respectively. The inverse, conjugate transpose and transpose operation is represented with the superscript $(\cdot)^{-1}$, $(\cdot)^{\mathrm{H}}$ and $(\cdot)^{\mathrm{T}}$, respectively. Cardinality of a set $\mathcal{X}$ is denoted with $|\mathcal{X}|$. The norm and the real part of a complex number is represented with $|\cdot|$ and $\Re\{\cdot\}$, respectively. Notation $\mathbb{C}^{M\times N}$ is a $M\times N$ matrix with elements in the complex field. $[\mathbf{a}]_n$ is the nth element of $\mathbf{a}$, and $(a)^+ \triangleq \max(0, a)$. Notation $\nabla_{\mathbf{x}} y(\mathbf{x})$ represent the gradient of $y(\cdot)$ with respect to $\mathbf{x}$.

II. SYSTEM ARCHITECTURE

We consider a downlink transmission in a mmWave based cloud (or centralize) radio access network (C-RAN) architecture, where all RRUs are connected to the edge cloud by the fronthaul links, as illustrated in Fig. 1. Each RRU is equipped with N transmit antennas. We use $\mathcal{K} = \{1, 2, \ldots, K\}$ to denote the set of all single antenna UEs, and $\mathcal{B} = \{1, 2, \ldots, B\}$ to denote the set of all RRUs. Further, the set of RRUs that serve kth UE is denoted by $\mathcal{B}_k$, such that $\mathcal{B}_k \subseteq \mathcal{B}$, $\forall k \in \mathcal{K}$. We assume JT-CoMP connectivity [7], where each user k receives a synchronous signal from its serving RRUs $\mathcal{B}_k$ $\forall k$. We assume that the network operates in a time-slotted manner, and the slots are normalized to an integer value, e.g., $t \in \{1, 2, \ldots\}$. Further, we assume that all RRUs use the same time-frequency resources for data transmission.

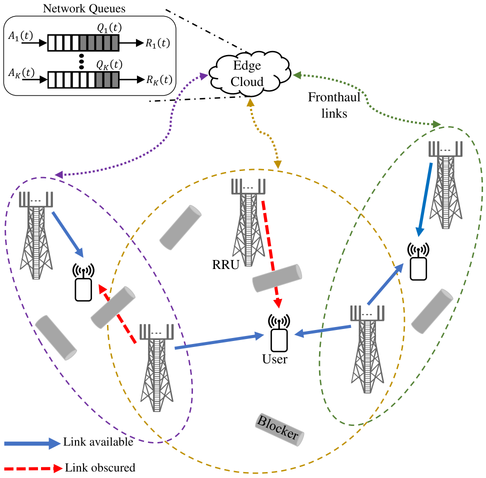

FIGURE 1. **C-RAN with network queues, *B* transmitters (RRUs), and *K* receivers (UEs) in the presence of randomly distributed blockers.**

Let $\mathbf{f}_{b,k}(t) \in \mathbb{C}^{N\times 1}$ denote the transmit beamforming vector from bth RRU to kth UE. Then, the received signal $y_k(t)$ at kth UE during time slot t can be expressed as

$$y_k(t) = \sum_{b\in\mathcal{B}_k} \mathbf{h}^{\mathrm{H}}_{b,k}(t)\mathbf{f}_{b,k}(t)d_k(t) + \sum_{u\in\mathcal{K}\setminus k}\ \sum_{g\in\mathcal{B}_u} \mathbf{h}^{\mathrm{H}}_{g,k}(t)\mathbf{f}_{g,u}(t)d_u(t) + w_k(t), \qquad (1)$$

where $\mathbf{h}_{b,k}(t) \in \mathbb{C}^{N\times 1}$ is the channel vector between RRU-UE pair (b, k). Notation $d_k(t)$ is data symbol associated with kth UE, and $w_k(t) \in \mathcal{CN}(0, \sigma_k^2)$ is circularly symmetric additive white Gaussian noise (AWGN). Moreover, we assume that data symbols are normalized and independent, i.e., $\mathbb{E}\{|d_k(t)|^2\} = 1$ and $\mathbb{E}\{d_k(t)d_u^*(t)\} = 0$ for all $k, u \in \mathcal{K}$. The received signal-to-interference-plus-noise ratio (SINR) of kth UE during time slot t can be expressed as

$$\Gamma_k(\mathbf{F}(t)) = \frac{\left|\sum_{b\in\mathcal{B}_k} \mathbf{h}^{\mathrm{H}}_{b,k}(t)\mathbf{f}_{b,k}(t)\right|^2}{\sigma_k^2 + \sum_{u\in\mathcal{K}\setminus k}\left|\sum_{g\in\mathcal{B}_u} \mathbf{h}^{\mathrm{H}}_{g,k}(t)\mathbf{f}_{g,u}(t)\right|^2}, \qquad (2)$$

where $\mathbf{F}(t) \triangleq [\mathbf{f}_{1,1}(t), \mathbf{f}_{1,2}(t), \ldots, \mathbf{f}_{B,K}(t)]$.

A. BLOCKAGE MODEL AND ACHIEVABLE RATE

In mmWave frequency band, the radio channel is spatially sparse due to low-scattering, reduced diffraction, and higher penetration and path losses [3], [4]. Hence, a mmWave communication link is inherently unreliable due to its susceptibility to blockages. The channel measurements in a typical mmWave outdoor scenarios have empirically shown that a link outage occurs with $20\% - 60\%$ probability [3], and may lead to a $10-$fold decrease in the achievable sum-rate performance [4]. Therefore, unless being addressed properly, the random link blockage appears as the main bottleneck hindering the full exploitation of the large available bandwidth at higher frequencies. Thus, to characterize the aforementioned uncertainties of the mmWave channel, we consider a probabilistic binary blockage model [21], [22]. Specifically, the radio channel $\mathbf{h}_{b,k}(t)$ between RRU-UE pair (b, k) can have one of two states, i.e., it is either fully-available or completely blocked. Furthermore, we consider link specific blockage, and the blocking of channel $\{\mathbf{h}_{b,k} = \mathbf{0}\}_{b\in\mathcal{B},k\in\mathcal{K}}$ for all t is independent. The methodology can easily be extended to more elaborate channel blocking models, e.g., by considering the distance and the spatio-temporal correlations [23]. In fact, this is an interesting topic for future extensions.

Similarly to [6], for improving system reliability and avoiding outage under the uncertainties of mmWave channel, we preemptively underestimate the achievable SINR assuming that a portion of CoMP links would be blocked during the downlink data transmission phase. This is specifically required in the mmWave communication because of dynamic blockages, which are, in general, not possible to track during the channel estimation phase. Let baseband processing unit (BBU) assume that each user k have at least $L_k(t)\in\big[1, |\mathcal{B}_k|\big]$ available links (i.e., unblocked RRUs). Then, we allow BBU to proactively model the lower-bound of achievable SINR over all possible subset combinations, e.g., by excluding the potentially blocked links, and allocate the pessimistic rate to users. As an example, let the set of RRUs that are used to serve kth UE with RRU indices $\mathcal{B}_k = \{1, 2, 3\}$. Then, with the assumption of at least $L_k(t) = 2$ available links, the serving set of unblocked RRUs available to kth UE can be any one of following combinations:

$$\widehat{\mathcal{B}}_k(L_k(t)) = \big\{\{1, 2\}, \{1, 3\}, \{2, 3\}, \{1, 2, 3\}\big\}. \tag{3}$$

Let $C(L_k(t))$ represent the cardinality of set $\widehat{\mathcal{B}}_k(L_k(t))$, defined as $C(L_k(t))=\sum_{l=L_k(t)}^{|\mathcal{B}_k|}\frac{|\mathcal{B}_k|!}{l!\,(|\mathcal{B}_k|-l)!}$. We use $\mathcal{B}_k^{\mathrm{c}}$ to denote the c-th subset of $\widehat{\mathcal{B}}_k(L_k(t))$, i.e., $\mathcal{B}_k^{\mathrm{c}} \in \widehat{\mathcal{B}}_k(L_k(t))$ such that cardinality $|\mathcal{B}_k^{\mathrm{c}}| \geq L_k(t)$ for all $c = 1, \ldots, C(L_k(t))$ during time slot t. Then, the received SINR of kth UE for $\mathcal{B}_k^{\mathrm{c}}$ (i.e., c-th subset) is obtained by excluding the potentially blocked RRUs in expression (2), and it can be expressed as

$$\Gamma_k(\mathbf{F}(t), \mathcal{B}_k^{\mathrm{c}}) = \frac{\Big|\sum\limits_{b\in\mathcal{B}_k^{\mathrm{c}}} \mathbf{h}_{b,k}^{\mathrm{H}}(t)\mathbf{f}_{b,k}(t)\Big|^2}{\sigma_k^2 + \sum\limits_{u\in\mathcal{K}\backslash k}\Big|\sum\limits_{g\in\mathcal{B}_u\backslash\mathcal{D}_k^{\mathrm{c}}} \mathbf{h}_{g,k}^{\mathrm{H}}(t)\mathbf{f}_{g,u}(t)\Big|^2}, \tag{4}$$

where $\mathcal{D}_k^{\mathrm{c}} = \mathcal{B}_k\backslash\mathcal{B}_k^{\mathrm{c}}$ denotes a subset of potentially blocked RRUs which are excluded from the interfering links to kth UE. Thus, the pessimistic achievable rate for kth UE during time slot t can be expressed as

$$r_k(t) = \log_2\big(1 + \gamma_k(t)\big), \tag{5}$$

where $\gamma_k(t) = \min\limits_{c}\big(\Gamma_k(\mathbf{F}(t), \mathcal{B}_k^{\mathrm{c}})\big), \forall k,\ c = 1, \ldots, C(L_k(t))$. In practice, the adverse channel condition and signaling overhead limits the maximum number of cooperating RRUs for each user (i.e., $\mathcal{B}_k\ \forall k$) [24]. Thus, the subset combinations $C(L_k(t))$ are fairly small for modestly sized systems [6].

It should be noted that each UE k still coherently receives the signal from all $\mathcal{B}_k$, unless RRUs are not available during the downlink data transmission phase due to random blockage. However, the actual RRUs available to serve kth UE cannot be known a priori in the dynamic blockage environment. Therefore, to maintain a reliable connectivity at each time slot, BBU preemptively underestimate the achievable SINR over all possible subset combinations $\mathcal{B}_k^{\mathrm{c}} \in \widehat{\mathcal{B}}_k(L_k(t))$, which is associated with the design parameter $L_k(t)\ \forall k$, as it will become clear in Section V.

B. NETWORK QUEUEING MODEL

Due to the stochastic nature of the mmWave channel, time-varying link qualities, and limited radio resources, there exists a possibility that the link between the RRU-UE pair is in poor conditions (i.e., in a deep fading state). Thus, scheduling such UEs will provide a little (if any) benefit. However, saving the network power and waiting for better channel conditions may lead to improved system performance [25]. Therefore, we assume that the BBU maintains a set of internal queues for storing network layer data of all UEs [19, Ch. 5]. Note that queue buffer provides a new degree-of-freedom to schedule the transmissions and flexibility to dynamically allocate resources over the fading channel states, while capturing the non-stationary evolution of data traffic per user. Let $Q_k(t)$ denote the current queue backlog of kth UE during time slot t, and $A_k(t)$ represents the amount of data that exogenously arrive to it, with the mean arrival rate of $\mathbb{E}[A_k(t)]=\lambda_k$. Then the dynamics of queue $Q_k(t)$ can be expressed as

$$Q_k(t+1) = \big[Q_k(t) - r_k(t) + A_k(t)\big]^+,\ \ \forall k \in \mathcal{K}, \tag{6}$$

where $r_k(t)$ is transmission rate defined in expression (5). Furthermore, let $\overline{Q}_k$ denote the time-averaged queue associated with kth UE, defined as

$$\overline{Q}_k \triangleq \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\big[Q_k(t)\big], \tag{7}$$

where the expectation $\mathbb{E}[\cdot]$ depends on the control policy, and is with respect to the random channel states and data arrivals.

According to the Little's law, the average delay is directly proportional to the average queue length $\overline{Q}_k$ [26, Ch. 1.4]. Hence, for kth UE, we can achieve the desired latency requirements by imposing a constraint on its queue length at each time slot. Here, we use a probabilistic constraint on the queue length, which is defined as

$$\Pr\left\{Q_k(t) \geq Q_k^{\text{th}}\right\} \leq \epsilon, \ \ \forall t, \tag{8}$$

where Q_k^{th} is the allowable queue backlogs for kth UE and $\epsilon \ll 1$ is the tolerable queue length violation probability.

C. PROBLEM FORMULATION

Our objective is to develop a power-efficient and reliable downlink transmission strategy for C-RAN based dynamic mmWave systems, while satisfying the user-specific latency requirements. Specifically, we consider a problem of time-average sum-power minimization for mmWave communication with random link blockages, subject to the maximum allowable queue length constraint for each UE, expressed as

$$\min_{\mathbf{F}(t), \gamma_k(t), \ \forall t} \ \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1}\left(\sum_{b\in\mathcal{B}}\sum_{k\in\mathcal{K}}\mathbb{E}\left[\|\mathbf{f}_{b,k}(t)\|^2\right]\right) \tag{9a}$$

$$\text{s.t.} \ \Pr\left\{Q_k(t) \geq Q_k^{\text{th}}\right\} \leq \epsilon, \ \ \forall k \in \mathcal{K}, \ \ \forall t \tag{9b}$$

$$\gamma_k(t) = \min_c\left(\Gamma_k(\mathbf{F}(t), \mathcal{B}_k^{\text{c}})\right), \quad \forall c = 1, \ldots, C(L_k(t)), \ \ \forall k \in \mathcal{K}, \ \ \forall t, \tag{9c}$$

where the function $\Gamma_k(\mathbf{F}(t), \mathcal{B}_k^{\text{c}})$ is defined in (4). The constraint (9b) ensures that the queue backlog of each user is less than Q_k^{th} at each time slot with the probability $(1-\epsilon)$, and thus ensures the desired probabilistic latency requirements. Note that for each user k, the constraint (9c) is a pessimistic estimate of achievable SINR. More specifically, for a given parameter $L_k(t)$, BBU models the SINR of kth UE over all possible subset combinations of potentially available RRUs $\mathcal{B}_k^{\text{c}}$ from the serving set $\widehat{\mathcal{B}}_k(L_k(t))$ (see Section II-A). Then, we allow BBU to proactively use the pessimistic estimate of SINR in order to allocate the downlink rate for the users such that transmission reliability is improved (i.e., to minimize the outage due to random link blockages that appear during the downlink transmission phase).

III. DYNAMIC CONTROL ALGORITHM

The problem (9) is intractable as it consists of a long-term time-average sum-power objective function (9a), non-linear probabilistic queue length constraint (9b), and a large number of coupled non-convex SINR expressions (9c). In this section, we handle the first two sources of intractability, and derive a dynamic control algorithm for (9) by using the Lyapunov optimization framework [19]. The proposed convex relaxations via FP techniques and the closed-form iterative algorithms are then provided in Section IV.

We start by *upper-bounding* the probabilistic queue length constraint (9b) as a time-average constraint using well-known Markov's inequality, i.e., $\Pr\{Q_k \geq Q_k^{\text{th}}\} \leq \mathbb{E}[Q_k]/Q_k^{\text{th}}, \ \forall k$ [26]. Thereby, problem (9) can be rewritten as

$$\min_{\mathbf{F}(t), \gamma_k(t), \ \forall t} \ \lim_{T\to\infty} \frac{1}{T}\sum_{t=0}^{T-1}\left(\sum_{b\in\mathcal{B}}\sum_{k\in\mathcal{K}}\mathbb{E}\left[\|\mathbf{f}_{b,k}(t)\|^2\right]\right) \tag{10a}$$

$$\text{s.t.} \ \lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[Q_k(t)\right] \leq \epsilon Q_k^{\text{th}}, \ \ \forall k \in \mathcal{K}, \ \ \forall t \tag{10b}$$

$$\gamma_k(t) \leq \Gamma_k(\mathbf{F}(t), \mathcal{B}_k^{\text{c}}), \quad \forall c = 1, \ldots, C(L_k(t)), \ \ \forall k \in \mathcal{K}, \ \ \forall t. \tag{10c}$$

Note that we have relaxed (9c) while writing constraint (10c), and both these constraints are equivalent at the optimality.

Now, we use the Lyapunov framework, specifically, a drift-plus-penalty method [19] to find a solution of problem (10). We enforce the long-term time-average constraint (10b) by transforming it into a queue stability problem [19, Ch. 5]. Specifically, a virtual queue associated with (10b) for each user k is introduced, and the stability of these virtual queues implies that the constraint (10b) is met.

Let $Z_k(t)$ be the virtual queue associated with (10b) for kth UE, and we update $Z_k(t)$ as

$$Z_k(t+1) = \left[Z_k(t) + Q_k(t+1) - \epsilon Q_k^{\text{th}}\right]^+, \ \ \forall k \in \mathcal{K}. \tag{11}$$

The expression (11) can be interpreted as a queue dynamics for kth UE with arrival rate $Q_k(t+1)$ and service rate ϵQ_k^{th}. It can be observed from (11) that if the queue length of a user is larger than the delay tolerance, the virtual queue will increase. Therefore, if the virtual queues $\{Z_k(t)\}_{k\in\mathcal{K}}$ are stable, then by using [19, Theorem 2.5] we can show that constraint (10b) is satisfied.

We now define Lyapunov function and its drift, which is used to stabilize queues $\{Z_k(t)\}_{k\in\mathcal{K}}$. For a compact representation, let $\Psi(t) = [Z_1(t), \ldots, Z_K(t), Q_1(t), \ldots, Q_K(t)]^{\text{T}}$ denote a collection of virtual and actual queues. Then we define following quadratic Lyapunov function [19, Ch. 3]:

$$\mathcal{L}(\Psi(t)) \triangleq \frac{1}{2}\sum_{k\in\mathcal{K}} Z_k(t)^2. \tag{12}$$

Intuitively, we can observe that if $\mathcal{L}(\Psi(t))$ is small, then all queues $\{Z_k(t)\}_{k\in\mathcal{K}}$ are small. Contrarily, if $\mathcal{L}(\Psi(t))$ is large then at least one of the queues is large. Thus, by minimizing a drift of $\mathcal{L}(\Psi(t))$ from one time slot to another, queues virtual $\{Z_k(t)\}_{k\in\mathcal{K}}$ can be stabilized, and thus, pushing the network queue backlog towards the desired requirements.

Then the Lyapunov drift [19, Ch. 5], which describes the change in network congestion between consecutive time slots, can be expressed as

$$\begin{aligned}\triangle(\Psi(t)) &= \mathbb{E}\left[\mathcal{L}(\Psi(t+1)) - \mathcal{L}(\Psi(t))|\Psi(t)\right] \\ &= \frac{1}{2}\mathbb{E}\left[\sum_{k\in\mathcal{K}}\left(Z_k(t+1)^2 - Z_k(t)^2\right)\middle|\Psi(t)\right].\end{aligned} \tag{13}$$

Next, by using expressions (6) and (11) in (13), an upper bound of drift $\triangle(\Psi(t))$ can be expressed as[1]

$$\begin{aligned}\triangle(\Psi(t)) \leq \zeta + \Phi(t) \ - \mathbb{E}\Big[\sum_{k\in\mathcal{K}}&(Q_k(t) + A_k(t) \\ &+ Z_k(t))r_k(t)\Big|\Psi(t)\Big],\end{aligned} \tag{14}$$

[1] To obtain (14), we have used the fact that $([a+b-c]^+)^2 \leq (a+b-c)^2$ for any $a \geq 0$, $b \geq 0$, and $c \geq 0$.

where ζ and $\Phi(t)$ are positive constants, and satisfy the following condition[2] for all time slots:

$$\zeta \geq \frac{1}{2}\mathbb{E}\Big[\sum_{k\in\mathcal{K}}\{A_k(t)^2 + r_k(t)^2\}\big|\Psi(t)\Big],$$

$$\Phi(t) = \sum_{k\in\mathcal{K}}\Big[\frac{1}{2}(\epsilon Q_k^{\text{th}})^2 + \frac{1}{2}Q_k(t)^2 + Z_k(t)Q_k(t) + \big(Q_k(t) + Z_k(t)\big)A_k(t)\Big].$$

Now we define following drift-plus-penalty function [19] for problem (10):

$$\triangle(\Psi(t)) + V\mathbb{E}\Big[\sum_{b\in\mathcal{B}}\sum_{k\in\mathcal{K}}\|\mathbf{f}_{b,k}(t)\|^2\big|\Psi(t)\Big], \tag{15}$$

where $V \geq 0$ is a trade-off parameter.[3] By using expression (14) in (15), and minimizing the upper bound of (15) subject to constraint (10c) at each time slot, we can stabilize queues $\{Z_k(t)\}_{k\in\mathcal{K}}$ and minimize the sum power objective function of problem (10). Thus, we utilize the concept of opportunistic minimization of an expectation [19, Ch. 1.8] to minimize the drift-plus-penalty function (15), and obtain a dynamic control algorithm as detailed in Algorithm 1.

Algorithm 1 Dynamic Control Algorithm for (9)

```
For a given time slot $t$, observe current queue backlogs $\{Q_k(t),\ Z_k(t)\}$ and solve following problem:
    $\min_{\mathbf{F}(t),\gamma_k(t)}\ V\sum_{b\in\mathcal{B}}\sum_{k\in\mathcal{K}}\|\mathbf{f}_{b,k}(t)\|^2 - \sum_{k\in\mathcal{K}}\big(Q_k(t) + A_k(t) + Z_k(t)\big)\log_2\big(1+\gamma_k(t)\big)$   (16a)
    s.t. $\gamma_k(t) \leq \Gamma_k(\mathbf{F}(t), \mathcal{B}_k^{\text{c}}),\ \forall c,\ \forall k \in \mathcal{K}.$   (16b)
Update queues $Q_k(t+1)$ and $Z_k(t+1)$ by using (6) and (11), respectively, for all $k \in \mathcal{K}$
Set $t = t + 1$, and go to step 1
```

We can observe from Algorithm 1 that the queue length of current unserved requests buffer will subsequently impact the resource assignment decision in the next slot due to evolving queue backlog. At each time slot of Algorithm 1, we need to solve problem (16) to find beamforming vectors. Therefore, we derive iterative algorithms for this in the next section.

IV. ITERATIVE ALGORITHMS FOR PROBLEM (16)

The problem (16) is intractable as-is, mainly due to coupled and non-convex SINR expressions (16b). In this section, we elaborate on finding the solution of problem (16) by using the FP quadratic transform techniques [20]. In addition, we also provide a low-complexity beamformer design via iterative evaluation of the closed-form KKT optimality conditions.

It is worth highlighting, in the FP techniques [20], [27], [28], all non-convex constraints are approximated with the sequence of convex subsets, and the underlying convex subproblem is then iteratively solved until the desired convergence of objective function. The FP quadratic transform based solutions have been widely studied in many applications, e.g., power control, energy efficiency, and multi-antenna interference coordination [20], [27], [28]. For example, the non-convex SINR relaxation via FP techniques is provided for downlink in [20, Section IV], [28, Section III] and for uplink in [27, Section V], assuming perfect CSI and no blockages. Further, all these algorithms are studied for the static case (i.e., the resource allocation problem for a given instance). In view of the prior works, there lacks a systemic approach for the design of beamforming vectors in the JT-CoMP scenarios, while accounting for the uncertainties of mmWave radio channel, in a time-average dynamic network, with the stringent user-specific latency and reliability requirements, and thus, motivating the current work.

A. SOLUTION VIA FP TECHNIQUES

We start by using the expression of $\Gamma_k(\mathbf{F}, \mathcal{B}_k^{\text{c}})$ (see (4)), and compactly rewrite (16b) as

$$\gamma_k(t) \leq \frac{|\mathbf{h}_k^{\text{cH}}(t)\mathbf{f}_k(t)|^2}{\sigma_k^2 + \sum_{u\in\mathcal{K}\backslash k}\big|\mathbf{h}_k^{\text{cH}}(t)\mathbf{f}_u(t)\big|^2}, \tag{17}$$

where $\mathbf{f}_k(t) \triangleq [\mathbb{1}_{\mathcal{B}_k}(1)\mathbf{f}_{1,k}^{\text{T}}(t), \ldots, \mathbb{1}_{\mathcal{B}_k}(B)\mathbf{f}_{B,k}^{\text{T}}(t)]^{\text{T}} \in \mathbb{C}^{|\mathcal{B}|N\times 1}$ and $\mathbf{h}_k^{\text{c}}(t) \triangleq [\mathbb{1}_{\mathcal{G}_k^{\text{c}}}(1)\mathbf{h}_{1,k}^{\text{T}}(t), \ldots, \mathbb{1}_{\mathcal{G}_k^{\text{c}}}(B)\mathbf{h}_{B,k}^{\text{T}}(t)]^{\text{T}} \in \mathbb{C}^{|\mathcal{B}|N\times 1}$ denotes the stacked beamformer and channel, respectively. The indicator function $\mathbb{1}_{\mathcal{G}_k^{\text{c}}}(b)$ and $\mathbb{1}_{\mathcal{B}_j}(b)$ are defined as

$$\mathbb{1}_{\mathcal{B}_k}(b) = \begin{cases} 1 & \text{if and only if} \quad b \in \mathcal{B}_k \\ 0 & \text{otherwise.} \end{cases}$$

$$\mathbb{1}_{\mathcal{G}_k^{\text{c}}}(b) = \begin{cases} 1 & \text{if and only if} \quad b \in \mathcal{B} \setminus \mathcal{D}_k^{\text{c}} \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathcal{G}_k^{\text{c}} = \mathcal{B}\backslash\mathcal{D}_k^{\text{c}}$ for all $c = 1, \ldots, C(L_k)$ and $k \in \mathcal{K}$. In this section, we omit time index t to simplify the notations.

Let us now examine the characteristic of the objective function in problem (16). We can observe that the right-hand-side (RHS) of (16b) is a typical function of multiple fractional parameters (i.e., SINR expressions (17)). Thus, problem (16) can be recast as a multi-ratio fractional problem. Motivated by the findings in [20], here we adopt the FP quadratic transform techniques, wherein the non-convex problem is recast as a sequence of convex subproblems, and then iteratively solved until the desired convergence of objective function. We extend the approaches [20] to take into consideration coherent multi-point transmission and provide a novel grouping of a multitude of potentially coupled and non-convex SINR conditions, that raise from the link blockage subset combinations of RRUs. We develop the following proposition based on the FP techniques [20, Theorem 1].

[2] We have assumed that second moments of arrival and transmission processes are bounded [19]. The derivation is omitted due to lack of space, and we refer the reader to [19] for the details.

[3] Note that V is a control parameter, reflecting the importance of objective function, i.e., higher values of V emphasize the minimization of the sum-power at the expense of linearly increasing the queue length, and vice versa.

Proposition 1: The fractional terms in RHS of (17)

$$\frac{|\mathbf{h}_k^{\mathrm{cH}}\mathbf{f}_k|^2}{\sigma_k^2 + \sum_{u\in\mathcal{K}\setminus k} \left|\mathbf{h}_k^{\mathrm{cH}}\mathbf{f}_u\right|^2}, \tag{18}$$

is equivalent to

$$2\Re\left\{\nu_{k,c}^{*}\mathbf{h}_k^{\mathrm{cH}}\mathbf{f}_k\right\} - \nu_{k,c}^{*}\left[\sigma_k^2 + \sum_{u\in\mathcal{K}\setminus k} \left|\mathbf{h}_k^{\mathrm{cH}}\mathbf{f}_u\right|^2\right]\nu_{k,c}, \tag{19}$$

when the auxiliary variable $\{\nu_{k,c}\}$ has the optimal value as

$$\nu_{k,c}^{(\star)} = \frac{\mathbf{h}_k^{\mathrm{cH}}\mathbf{f}_k}{\sigma_k^2 + \sum_{u\in\mathcal{K}\setminus k} \left|\mathbf{h}_k^{\mathrm{cH}}\mathbf{f}_u\right|^2}, \tag{20}$$

for all $c = 1, \ldots, C(L_k)$ and $k \in \mathcal{K}$. The Proposition 1 can easily be proved by following the steps in [20, Section IV].

Thereby using Proposition 1, we can obtain a solution for problem (16) by iteratively solving a sequence of convex subproblems [20]. For example, the convex subproblem for i-th iteration along with dual variables can be expressed as

$$\min_{\mathbf{F},\gamma_k} V\sum_{k\in\mathcal{K}} \|\mathbf{f}_k\|^2 - \sum_{k\in\mathcal{K}} (Q_k + A_k + Z_k)\log_2(1+\gamma_k) \tag{21a}$$

$$\begin{aligned}\text{s.t. } e_{k,c} : \gamma_k \le\; & 2\Re\left\{\nu_{k,c}^{*(i-1)}\mathbf{h}_k^{\mathrm{cH}}\mathbf{f}_k\right\} \\ & - \nu_{k,c}^{*(i-1)}\left[\sigma_k^2 + \sum_{u\in\mathcal{K}\setminus k} \left|\mathbf{h}_k^{\mathrm{cH}}\mathbf{f}_u\right|^2\right]\nu_{k,c}^{(i-1)}, \;\; \forall c, \;\; \forall k \in \mathcal{K},\end{aligned} \tag{21b}$$

where $\{e_{k,c}\}$ are non-negative Lagrangian multipliers associated with constraints (21b). Note that (21) provides an approximate solution for (16) in the vicinity of a fixed operating point $\{\nu_{k,c}^{(i-1)}\}$. Specifically, for fixed auxiliary variables, we first optimize the primal optimization variables, and then update auxiliary variables with the current solution using expression (20). Hence, by iteratively solving (21) while updating the auxiliary variables $\{\nu_{k,c}^{(i)}\}$ with current solution, we can find a best solution for (16), using existing convex optimization toolboxes, such as CVX [29]. The beamformer design with the proposed FP relaxations has been summarized in Algorithm 2.

Algorithm 2 FP based Algorithm for (16)

```
Set $i = 1$ and initialize with a feasible $\left\{\nu_{k,c}^{(0)}\right\}, \forall c, \forall k$
repeat
    Solve (21) with $\left\{\mathbf{v}_{k,c}^{(i-1)}\right\}$ and denote the local
    solution as $\left\{\mathbf{f}_k^{(i)}, \gamma_k^{(i)}\right\}$
    Obtain $\nu_{k,c}^{(i)}$ using (20) with updated $\left\{\mathbf{f}_k^{(i)}\right\}$
    Set $i = i + 1$
until convergence or for fixed number of iterations
```

B. SOLUTION VIA KKT CONDITIONS

We can observe that for fixed auxiliary variables $\{\nu_{k,c}\}$, each subproblem (21) is a convex problem with respect to variables $\{\mathbf{f}_k, \gamma_k\}$. Thus, we can efficiently obtain the solution by the Lagrangian multiplier method. Here, we tackle (21) by iteratively solving a system of KKT optimality conditions [30, Ch. 5.5], and, in general, it admits a closed-form solution that does not rely on generic (or tailored) convex solvers. The Lagrangian $\mathfrak{L}_{\mathrm{FP}}(\mathbf{F}, \gamma_k, \nu_{k,c}, e_{k,c})$ of problem (21) is given in (22), as shown at the bottom of the next page. The stationary conditions for kth UE is obtained by differentiating (22) with respect to associated primal optimization variables $\{\mathbf{f}_k, \gamma_k\}$ for all $k \in \mathcal{K}$ (refer to [30, Ch. 5.5.3] for details). Thus, the stationary conditions of each user k for problem (21) can be expressed as

$$\nabla_{\gamma_k} : \sum_{c=1}^{C(L_k)} e_{k,c} = \frac{Q_k + A_k + Z_k}{1+\gamma_k}, \tag{23a}$$

$$\begin{aligned}\nabla_{\mathbf{f}_k} : & \sum_{c=1}^{C(L_k)} e_{k,c}\left(\nu_{k,c}^{*(i-1)}\mathbf{h}_k^{\mathrm{cH}}\right) \\ & = \mathbf{f}_k^{\mathrm{H}}\left(V\mathbf{1} + \sum_{u\in\mathcal{K}\setminus k}\sum_{c=1}^{C(L_u)} e_{u,c}\nu_{u,c}^{*(i-1)}\left(\mathbf{h}_u^{\mathrm{c}}\mathbf{h}_u^{\mathrm{cH}}\right)\nu_{u,c}^{(i-1)}\right).\end{aligned} \tag{23b}$$

In addition to (23) and primal-dual feasibility constraints, the KKT conditions also include the complementary slackness as

$$\begin{aligned} e_{k,c} \ge 0; \quad e_{k,c}\Bigg\{&\gamma_k - 2\Re\left\{\nu_{k,c}^{*(i-1)}\mathbf{h}_k^{\mathrm{cH}}\mathbf{f}_k\right\} + \nu_{k,c}^{*(i-1)}\Bigg[\sigma_k^2 \\ &+ \sum_{u\in\mathcal{K}\setminus k}\left|\mathbf{h}_k^{\mathrm{cH}}\mathbf{f}_u\right|^2\Bigg]\nu_{k,c}^{(i-1)}\Bigg\} = 0, \;\; \forall c, \;\; \forall k \in \mathcal{K}.\end{aligned} \tag{24}$$

Note that the user-specific SINR constraint (21b) is mutually coupled and interdependent over the serving set combinations (see Section II-A). Hence, obtaining a closed-form solution for the associated Lagrangian multipliers $\{e_{k,c}\}$ in expression (23) is considerably more difficult than the case with single and non-coupled QoS constraint per user [7], [28]. Thus, to overcome this challenge, we resort to the subgradient approach, where all non-negative Lagrangian multipliers $\{e_{k,c}\}$ are iteratively solved using the subgradient method [31]. Furthermore, to avoid separate updates for inner and outer loops, and hence, to improve the convergence speed, the fixed operating point $\{\nu_{k,c}^{(i)}\}$ is also heuristically updated, in each iteration, along with primal variables and associated Lagrange multipliers. Thus, in general, the monotonic convergence can not be guaranteed, and it may not necessarily converge to the same solution as Algorithm 2. It is shown by numerical examples in Section VI that this still provides good performance with a fairly small number of approximation point updates. The closed-form steps in the proposed iterative algorithm are:

$$\begin{aligned}\mathbf{f}_k^{(i)\mathrm{H}} = & \sum_{c=1}^{C(L_k)} e_{k,c}^{(i-1)}\left(\nu_{k,c}^{*(i-1)}\mathbf{h}_k^{\mathrm{cH}}\right) \\ & \times\left\{V\mathbf{1} + \sum_{u\in\mathcal{K}\setminus k}\sum_{c=1}^{C(L_u)} e_{u,c}^{(i-1)}\nu_{u,c}^{*(i-1)}\left(\mathbf{h}_u^{\mathrm{c}}\mathbf{h}_u^{\mathrm{cH}}\right)\nu_{u,c}^{(i-1)}\right\}^{-1},\end{aligned} \tag{25a}$$

$$\gamma_k^{(i)} = \frac{Q_k + A_k + Z_k}{\sum_{c=1}^{C(L_k)} e_{k,c}^{(i-1)}} - 1, \tag{25b}$$

$$\Gamma_k^{(i)}(\mathbf{F}, \mathcal{B}_k^{\mathrm{c}}) = \frac{|\mathbf{h}_k^{\mathrm{cH}} \mathbf{f}_k^{(i)}|^2}{\sigma_k^2 + \sum_{u \in \mathcal{K}\setminus k} \left|\mathbf{h}_k^{\mathrm{cH}} \mathbf{f}_u^{(i)}\right|^2}, \tag{25c}$$

$$e_{k,c}^{(i)} = \left(e_{k,c}^{(i-1)} + \beta_e\big[\gamma_k^{(i)} - \text{expression (19)}\big]\right)^+, \tag{25d}$$

$$\nu_{k,c}^{(i)} = \frac{\mathbf{h}_k^{\mathrm{cH}} \mathbf{f}_k^{(i)}}{\sigma_k^2 + \sum_{u \in \mathcal{K}\setminus k} \left|\mathbf{h}_k^{\mathrm{cH}} \mathbf{f}_u^{(i)}\right|^2}. \tag{25e}$$

where β_e is small positive step-size.[4] In expression (25d), the dual variables $\{e_{k,c}\}$ are iteratively updated based on the violation of SINR constraint, e.g., the complementary slackness conditions (24), using the subgradient method [31]. The proposed beamformer design by solving a system of closed-form KKT expressions is summarized in Algorithm 3.

Algorithm 3 KKT based Iterative Algorithm for (21)

```
Set $i = 1$ and initialize $\{\nu_{k,c}^{(0)}\}$, $\forall c, \ \forall k$
repeat
    Solve $\mathbf{f}_k^{(i)}$ from (25a) with $\{\nu_{k,c}^{(i-1)}, e_{k,c}^{(i-1)}\}$
    Obtain $\gamma_k^{(i)}$ from (25b)
    Calculate $\Gamma_k^{(i)}(\mathbf{F}, \mathcal{B}_k^{\mathrm{c}})$ from (25c) with $\{\mathbf{f}_k^{(i)}\}$
    Update $e_{k,c}^{(i)}$ using (25d)
    Solve $\nu_{k,c}^{(i)}$ from (25e) with updated $\{\mathbf{f}_k^{(i)}\}$
    Set $i = i + 1$
until convergence or for fixed iterations
```

C. INITIALIZATION AND COMPLEXITY ANALYSIS

1) FEASIBLE INITIALIZATION

In the FP methods, the non-convex constraints (16b) are approximated with the sequence of convex subsets, and then iteratively solved until the convergence of objective function to a stationary point solution. Thus, it is important to initialize the proposed iterative algorithms with a feasible starting point, as it impacts the problem feasibility and the rate of convergence [20], [27]. To this end, one possible option for a feasible $\{\mathbf{f}_k^{(0)}\}$ is to use any randomly generated beamforming vector. Then, compute the lower bound of achievable SINR from (4), i.e., $\gamma_k^{(0)} = \min_c\big(\Gamma_k(\mathcal{B}_k^{\mathrm{c}})\big)$ for all $c = 1, 2, \ldots, C(L_k), \ k \in \mathcal{K}$. Furthermore, with the feasible $\{\mathbf{f}_k^{(0)}\}$, the initial values of the auxiliary variables $\{\nu_{k,c}^{(0)}\}$ can easily be computed using (20). The non-negative Lagrangian multiplier $\{e_{k,c}^{(0)}\}$ in Algorithm 3 are initialized such that $\sum_{c=1}^{C(L_k)} e_{k,c}^{(0)} > 0$, e.g., at least one of the coupled SINR constraint is active for each user (see (23a) and (25b) for more details). It is worth noting that the initialization of the iterative algorithms with different feasible initial values $\big\{\mathbf{f}_k^{(0)}, \gamma_k^{(0)}, \nu_{k,c}^{(0)}, e_{k,c}^{(0)}\big\}$, in general, does not impact the local solution of problem (16), provided a sufficient number of iterations [30]. For detail on the convergence and the stationary point solution, we refer the readers to [20], [27].

2) COMPUTATIONAL COMPLEXITY ANALYSIS

The approximated convex subproblem (21) can be solved in a generic convex optimization solver, i.e., as a sequence of second-order cone programs (SOCP) [32]. The interior points methods are generally adopted to efficiently solve the SOCP formulations, wherein, the computational complexity of each iteration scales with the length of system wide joint beamforming vectors ($|\mathcal{B}|N$) and the number of constraints [32]. In this case, it can be shown that solving each subproblem (21) requires $\mathcal{O}\big((|\mathcal{B}|N)^{3.5}\big)$ arithmetic operations. The computational complexity of the iterative algorithms, e.g., by solving a system of closed-form KKT optimality conditions, is mainly dominated by expression (25a) for each subproblem (21). We can observe that expression (25a) consists of matrix multiplications and inverse operations, and each iteration requires $\mathcal{O}\big((|\mathcal{B}_k|N)^{2.37}\big)$ arithmetic operations using, e.g., Coppersmith–Winograd algorithm [33], [30, Appendix C]. Thus, algorithms based on iterative evaluation of closed-form KKT optimality conditions provide relatively lower complexity.[5] compared to the joint beamformer optimization across all RRUs, and does not require any convex solver. As an example, let $|\mathcal{B}| = |\mathcal{B}_k| = 4$ and $N_t = \{4, 16\}$, Algorithm 3 resutls in the complexity reduction by $\{95.6\%, 99\%\}$, and thus, provides a solution for the practical implementations.

V. DYNAMIC SERVING SUBSET SELECTION

We assume that the blockers are randomly distributed and independent for each time slot. Further, the position of each blocker and/or blockage event can not be known during the downlink data transmission phase. Therefore, to improve system reliability under these uncertainties of mmWave radio channel, we preemptively underestimate

$$\mathfrak{L}_{\mathrm{FP}}(\mathbf{F}, \gamma_k, \nu_{k,c}, e_{k,c}) = \sum_{k=1}^{K}\Bigg[V\|\mathbf{f}_k\|^2 - (Q_k + A_k + Z_k)\log_2(1+\gamma_k) + \sum_{c=1}^{C(L_k)} e_{k,c}\gamma_k \\ -2\sum_{c=1}^{C(L_k)} e_{k,c}\Re\big\{\nu_{k,c}^{*(i-1)}\mathbf{h}_k^{\mathrm{cH}}\mathbf{f}_k\big\} + \sum_{c=1}^{C(L_k)} e_{k,c}|\nu_{k,c}^{(i-1)}|^2\sigma_k^2 + \sum_{u\in\mathcal{K}\setminus k}\sum_{c=1}^{C(L_u)} e_{u,c}\nu_{u,c}^{*(i-1)}|\mathbf{h}_u^{\mathrm{cH}}\mathbf{f}_k|^2\nu_{u,c}^{(i-1)}\Bigg]. \tag{22}$$

[4] The step size depends on the system model, as it directly affects the convergence rate and controls the oscillation in the objective function [31].

[5] As an alternative implementation, the matrix inversion in (25a) can be replaced with the best response framework [7, Section IV] [34], which efficiently parallelizes the beamformer updates across the RRU antennas.

the achievable rate of each user, assuming that a portion of CoMP links would be blocked during the data transmission phase (see Section II-A). Let BBU assume that each user k have at least $L_k(t)$ available links (i.e., unblocked RRUs). Then we allow BBU to proactively model the pessimistic estimate of SINR over all possible subset combinations, e.g., by excluding the potentially blocked links, and allocate the rate to users such that transmission reliability is improved (for more details see Section II-A). Hence, the user-specific CoMP subset combinations $L_k(t) \in [1,\ |\mathcal{B}_k|]\ \forall k$ is a design parameter, which can be tuned for each time slot t, e.g., based on available queue backlogs and channel information, to achieve desired rate, reliability, and user-specific latency requirements.

Let $\varrho_k(t) \in [0, 1]$ denote the blockage probability of kth UE during time slot t. Then, for a fixed subset combinations $L_k(t)$, the success probability of kth UE can be approximated as (refer to [6, Section III] for details)

$$p_k\big(L_k(t)\big) = \sum_{l=0}^{|\mathcal{B}_k|-L_k(t)} \binom{|\mathcal{B}_k|}{l} \big(1-\varrho_k(t)\big)^{|\mathcal{B}_k|-l} \big(\varrho_k(t)\big)^{l}. \quad (26)$$

Since all users are independent, the outage probability of kth UE can be expressed as

$$\mathrm{P}_k^{\mathrm{out}}\big(L_k(t)\big) = 1 - p_k\big(L_k(t)\big),\ \ \forall k \in \mathcal{K}. \quad (27)$$

From expression (27), we can observe that the outage $\mathrm{P}_k^{\mathrm{out}}\big(L_k(t)\big)$ is a monotonically increasing function of parameter $L_k(t)$. As an example, we can improve the system reliability and avoid the outage by preemptively assuming that a significant portion of the available CoMP RRUs (i.e., $|\mathcal{B}_k| - L_k(t),\ \forall k$) are potentially blocked. However, this pessimistic assumption on available links may lead to a lower SINR estimate (see Section II-A), and hence, a lower rate to each user (5). Thus, to ensure the user-specific latency requirements (9b), the network consumes more power, and attempt to increase the instantaneous achievable rate over the subset of potentially available RRUs, $\mathcal{B}_k^{\mathrm{c}}\ \forall k \in \mathcal{K}$.

Conversely, a less pessimistic assumption on subset size (i.e., a higher value of $L_k(t)\ \forall k$) can provide higher instantaneous SINR (4), but it can be more susceptible to the outage, and results in less stable connectivity. Moreover, these outage events will eventually increase the queue backlogs (6), i.e., due to unsuccessful downlink data transmissions. Thus, to guarantee the desired average latency requirements (9b), the network consumes more power, and tries to increase the achievable rates in the following transmit intervals. Clearly, there is a trade-off between reliable connectivity and sum-power performance, while ensuring the latency requirements.

Therefore, for each time slot t, first we need to choose parameter $L_k(t) \in [1,\ |\mathcal{B}_k|]$, and then solve problem (16) over a given subset combinations $C(L_k(t))$ for all $k \in \mathcal{K}$. Specifically, the parameter $L_k(t)$, such that the solution of problem (16) satisfy (9b) with the minimum sum-power is of our interest. Thus, we can observe that the constraint (9b) can be met with minimum sum-power, if the success probability of each user k satisfy $p_k\big(L_k(t)\big) \geq 1-\epsilon,\ \forall k \in \mathcal{K}$. Hence, the parameter $L_k(t)$ for kth UE during time slot t can be computed by solving following:

$$\min_{L_k(t)}\ p_k\big(L_k(t)\big) \quad (28a)$$

$$\text{s.t.}\ \ p_k\big(L_k(t)\big) \geq 1-\epsilon,\ \ \forall L_k(t) \in \big[1,\ |\mathcal{B}_k|\big]. \quad (28b)$$

It should be noted that the blockage probability $\varrho_k(t)\ \forall k$ is still an unknown parameter, and thus hinders solving (28). However, an approximation of blockage $\widetilde{\varrho}_k(t)\ \forall k$ can be obtained by, e.g., exploiting the available queue backlog information at the BBU. For example, in the considered C-RAN architecture, the centralized BBU is aware of the instantaneous arrivals, current queue backlogs, and channel information, for the design of beamforming vectors. Thus, user k during time slot t can be in outage, if the assigned downlink rates $r_k(t) \neq 0$ and the queue backlogs grow as $Q_k(t+1) = [Q_k(t) + A_k(t)]^+,\ \forall k \in \mathcal{K}$ (see expression (6) and (31) for more details). Therefore, the outage of kth UE during time slot t can be approximated as

$$\mathbb{1}_{\mathcal{P}_k}(t) = \begin{cases} 1 & \text{if}\ \big\{r_k(t) \neq 0\big\} \bigcap \\ & \qquad \big\{Q_k(t+1) = [Q_k(t) + A_k(t)]^+\big\}, \\ 0 & \text{otherwise.} \end{cases}$$

where $\mathbb{1}_{\mathcal{P}_k}(t)$ is indicator function. Alternatively, the outage event can be (more accurately) computed based on UE acknowledgments. However, in the presence of random link blockages, the mmWave feedback links are inherently unreliable, and hence, results in overestimation and increased delays [17], [18]. In fact, this is an interesting topic for future extensions. Thus, at each time slot t, BBU exploits the available channel and queue backlogs information of each user k to compute the approximated blockage as

$$\widetilde{\varrho}_k(t) = \frac{1}{\delta_k} \sum_{i=t-\delta_k}^{t-1} \mathbb{1}_{\mathcal{P}_k}(i), \quad (29)$$

where δ_k is maximum averaging length. Hence, using the estimated time-averaged blocking, the BBU first computes the adequate size of the subset combinations from (28), and then solves problem (16) to obtain the beamforming vectors.

VI. SIMULATION RESULTS

This section provides numerical examples to quantify the performance advantage of the proposed iterative algorithms. We consider a mmWave based donwlink transmission with UEs $K = 4$, RRUs $B = 4$, and each RRU is equipped with a uniform linear array (ULA) of $N = 16$ antennas. Further, RRUs are placed in a 50×50 meters square layout (resembling, e.g., a factory-type IIoT setup), and are connected to a common BBU in the edge cloud. All single antenna UEs are randomly dropped within the coverage region, thus each UE has a different path-gain and angle with the RRUs.

The mmWave channel $\mathbf{h}_{b,k}(t)$ between a RRU-UE pair (b, k) is based on sparse geometric model [35], and defined as

$$\mathbf{h}_{b,k}(t) = \sqrt{\frac{N}{M}} \sum_{m=1}^{M} \omega_{b,k}(t) d_{b,k}^{-\psi_m(t)}(t) \mathbf{a}_T^{\mathrm{H}}(\phi_{b,k}^m(t)), \quad (30)$$

where M is the number of independent paths, which we set as $M = 3$, and $\omega_{b,k}(t)$ is random complex gain with zero mean and unit variance. The distance between RRU-UE pair is represented with $d_{b,k}$, and the notation ψ_m denote a random path-loss exponent. In the simulation, we consider $\psi_m(t) \in [2, 6]$, $\forall(m, t)$. The array response vector for ULA is represented with $\mathbf{a}_T(\phi_{b,k}^m(t)) \in \mathbb{C}^{N\times 1}$, and relative to the boresight of the RRU antenna array, the angle-of-departure (AoD) for each path is uniformly distributed, i.e., $\phi_{b,k}^m(t) \in [-\pi/2, \pi/2]$, $\forall(m, t)$. For simplicity, we assume a probabilistic binary blockage model [21], [22], where the radio channel between RRU-UE pair during downlink transmission phase, is either fully available, i.e., as in expression (30) or completely blocked, i.e., $\{\mathbf{h}_{b,k}(t) = \mathbf{0}\}$, with the link blockage probability of $q \in [0, 1]$ for all $b \in \mathcal{B}$ and $k \in \mathcal{K}$.

Recall that to improve the communication reliability and avoid outage under the uncertainties of mmWave radio channel, we use parameter $L_k(\leq |\mathcal{B}_k|)$ in problem (9), and proactively model the SINR over the link blockage combinations (see Section II-A). For simplicity, but without loss of generality, we assume identical parameters for each user k, i.e., serving set $|\mathcal{B}_k| = 4$, $L_k = L$, data arrival $A_k \sim \mathrm{Pois}(\lambda)$ with $\lambda = 3.5$ bits/slot and maximum queue backlogs $Q_k^{\mathrm{th}} = 5$ bits with tolerable violation probability $\epsilon = 0.1$ in problem (9) [1]. In the simulations, we set the frequency $f_c = 28$ GHz, and the step size $\beta_e = 0.01$ in expression (25d).

The outage event occurs if the instantaneous transmit rate $r_k(t)$ exceeds the supported rate[6] $c_k(t)$ for all $k \in \mathcal{K}$. Then, the queue dynamics $Q_k(t)$ in (6) can be expressed as

$$Q_k(t+1) = \left[Q_k(t) - r_k(t)\mathbb{1}_{\{r_k(t)\leq c_k(t)\}} + A_k(t)\right]^+, \ \forall k, \quad (31)$$

where $\mathbb{1}_{\{\cdot\}}$ is an indicator function. Specifically, expression (31) implies that queue backlogs also increase with each unsuccessful downlink transmission due to random blockages. In our study, we will make use of subset size L to analyze the network performance. For the baseline methods, we consider CB (i.e., $|\mathcal{B}_k| = 1, \ \forall k$) [10] and full-JT (i.e., $L_k = |\mathcal{B}_k|, \ \forall k$) [7] based downlink beamformer designs.

A. CONVERGENCE ANALYSIS

In Fig. 2, we examine the convergence behavior of proposed iterative algorithms for given randomly generated channel realizations. For simplicity but without loss of generality, we set parameters $V=1$, $L=3$, and $q=0$. Note that the solution of Algorithms 2 is obtained directly by using the convex optimization toolbox, SeDuMi [29]. In contrast, Algorithm 3 is solved from a system of closed-form KKT optimality conditions in an iterative manner (see (25)). We can observe that, with the considered parameter settings, both algorithms converge to the same solution with a fairly small number of iterations. It is worth highlighting, in general, the convergence of Algorithm 3 cannot be guaranteed to be monotonic due to only a single subgradient updates (25d) in each iteration along with other variables (see Algorithm 3). We refer the reader to [31] on the convergence properties of the subgradient approach with different step size rules. To summarize, the Algorithm 2 provides monotonic and faster convergence in terms of required approximation point updates. On contrary, Algorithm 3 achieves comparable performance with a notable reduction in the per-iteration computational complexity and does not require any generic (or tailored) convex solvers, which can be useful for, e.g., hardware constrained devices with limited processing capabilities.

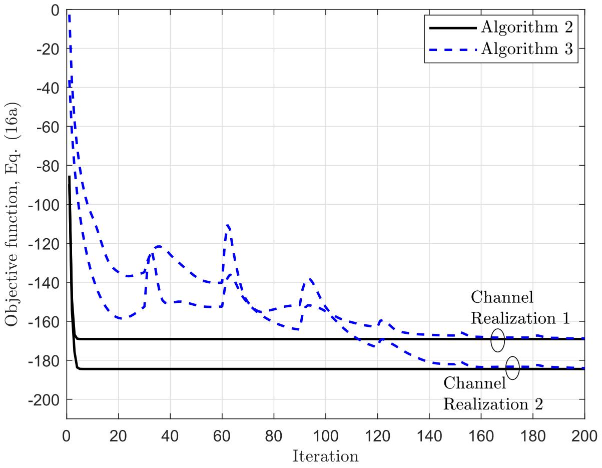

FIGURE 2. Convergence performance of Algorithm 2 and Algorithm 3.

B. IMPACT OF PARAMETER V

First, in Fig. 3, we illustrate the latency performance with trade-off parameter $V = 1$ and blockage $q = 10\%$. The result shows that our proposed and baseline methods satisfy the maximum queue backlogs of each user k (i.e., $Q_k^{\mathrm{th}} = 5$ bits) within the allowable queue tolerance level $\epsilon = 0.1$. Thus, problem (9) is feasible, and the proposed convex relaxations still allow to achieve the desired user-specific latency requirements (i.e., constraint (9b) is met). However, our proposed beamformer designs, i.e., by considering a pessimistic estimate of rates over the subset combinations of potentially blocked CoMP RRUs, substantially improve the average sum-power performance, while ensuring the same latency requirements, as shown in Fig. 4. As an example, for the parameter $V = 1$ and $L = 2$, our proposed method improves the average sum-power performance by 8 dBm and 18 dBm

[6] For a give time slot t, let $\mathcal{S}_k(t) = \{\gamma_k^\star(t), \mathbf{f}_k^\star(t)\}_{k\in\mathcal{K}}$ denote solution of problem (9). Then, for each UE k, the transmission rate is given by $r_k(t) = \log_2(1+\gamma_k^\star(t))$. However, the actual supported rate (i.e., link capacity) for kth UE depends on the obtained beamformers $\{\mathbf{f}_k^\star(t)\}_{k\in\mathcal{K}}$ and current channel state $\{\mathbf{h}_{b,k}(t)\}_{b\in\mathcal{B},k\in\mathcal{K}}$. However, channel can not be exactly known to the BBU during data transmission phase due to random blockages. Thus, the supported rate can be calculated by using the actual SINR values (2), i.e., $c_k(t) = \log_2(1+\Gamma_k(\mathbf{F}^\star(t)))$, $\forall k \in \mathcal{K}$, and these supported link rates are unknown to the BBU.

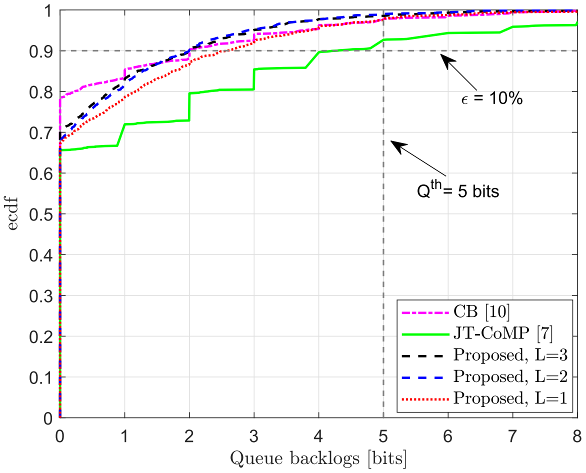

FIGURE 3. Queue backlogs with $V = 1$ and blockage of 10%.

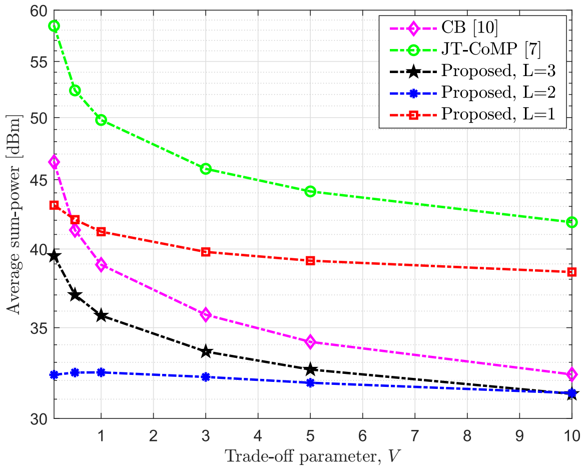

FIGURE 4. Average sum-power with increasing V and blockage of 10%.

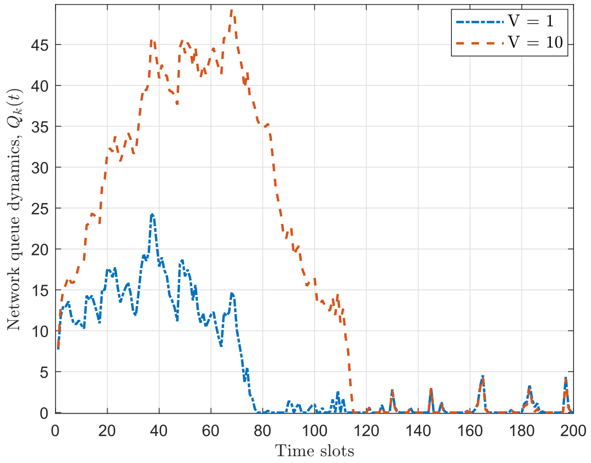

FIGURE 5. Network queue dynamics with blockage of 10%.

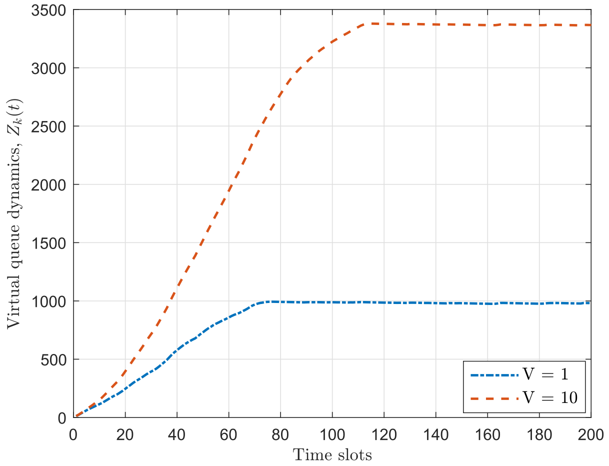

FIGURE 6. Virtual queue dynamics with blockage of 10%.

compared to baseline CB and full-JT, respectively. Hence, the proposed methods significantly outperform the conventional full-JT [7] and CB [10] based downlink beamformer design, and thus provides power-efficient and latency-constrained highly-reliable mmWave communication.

Further, it can be observed from Fig. 4 that for a fixed queue length constraint, the average sum-power decreases with the increase in the value of parameter V. This behavior is expected, since higher values of V linearly emphasize the minimization of the sum-power objective over the queue length, until the queue backlogs become substantially larger than the sum-power objective values (see (16a) for details).

Fig. 5 and Fig. 6 shows the evolution of the network queues $\{Q_k(t)\}$ and the associated virtual queues $\{Z_k(t)\}$ over time with the blockage $q = 10\%$ and the parameter $L = 3$, for user $k = 1$. Note that, we can observe similar behavior for all other users, but these are not included due to space limitations. It can be concluded from Fig. 5 that the queue backlogs increases differently for different values of trade-off parameters V, until it saturates and reaches a certain value (e.g., for $V = 10$ around $t = 115$ time slots), and then it oscillates, so as the constraint $\Pr\{Q_k(t) \geq Q_k^{\text{th}}\} \leq \epsilon$, is strictly ensured, i.e., to achieve the average user-specific latency requirements. This is mainly because of the negative drift property of the Lyapunov function [19, Ch. 4.4]. Thus, the stability of associated virtual queues $\{Z_k(t)\}$, i.e., as in Fig. 6, ensures that the network queues are bounded, and achieves the desired queue backlogs performance per user, i.e., constraint (9b) is satisfied.

C. DYNAMIC SELECTION OF PARAMETER L

Next, we investigate the dynamic selection of the serving set size, i.e., parameter $L(t)$ for each time slot t, which is obtained by solving (28). To do that, first in Fig. 7, we show the time-averaged blockage computed from expression (29) with the maximum averaging length $\delta_k = \min\{\tau, (t-1)\}$, and we set $\tau = 50$. It can be concluded that the estimate of the blockage in (29) closely matches with actual blockage probability, with a small number of random channel realizations. Thus, expression (29) provides a fair approximation, and the

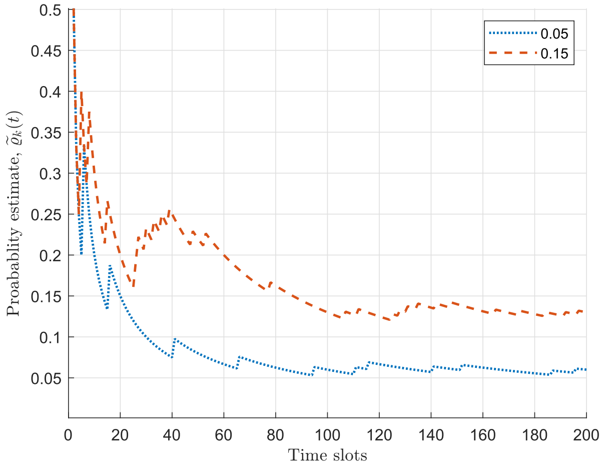

FIGURE 7. Approximation of blockage probability using (29).

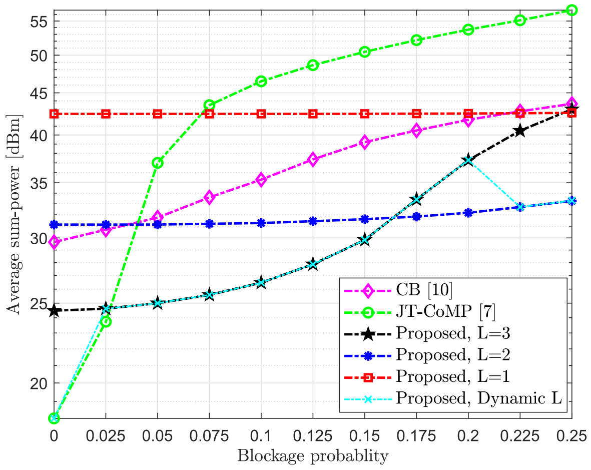

FIGURE 9. Average sum-power with increasing blockage and $V = 1$.

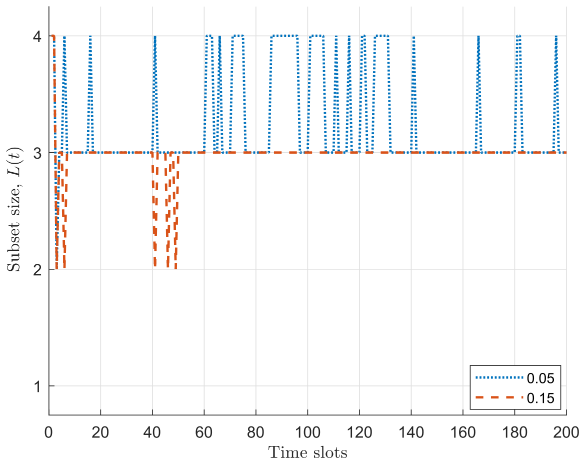

FIGURE 8. Dynamic selection of subset size $L(t)$ by solving (28).

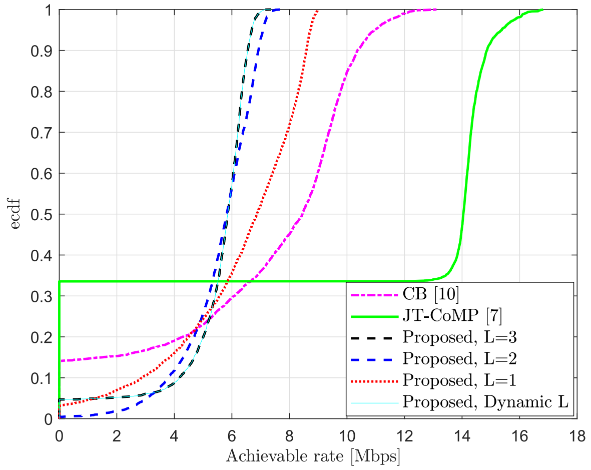

FIGURE 10. Downlink user-rate with $V = 1$ and blockage of 10%.

resulting gap is mainly due to unpredictable blockage events and uncertainties in the mmWave radio channel.

Further, it can be observed from Fig. 8 that the instantaneous choice of serving set size, i.e., to meet the success probability (28b), mainly depends on the accuracy of estimated blockage. Furthermore, the oscillations in Fig. 8 is due to limited (and discrete) choices of parameter $L_k(t) \in [1, |\mathcal{B}_k|], \forall k$, and possibly the minimum subset size satisfying (28b) can be in between the integer values. It is worth highlighting that the problem (28) aims at finding serving set size, such that solving problem (16) satisfy the (9b) with minimum sum-power. Therefore, large values of L may result in lower sum-power, but it may lead to higher outage, and thus it dynamically switches to a more pessimistic SINR estimate, i.e., to lower values of parameter L in the following time slot, to meet the average latency requirements. Hence, there is a trade-off between average sum-power, achievable rate, and reliability, as will be demonstrated in the following.

Fig. 9 illustrates the impact increasing blockage probability on constraint (9c) and sum-power performance. For example, the worst-case pessimistic assumption on available links (i.e., $L = 1$) leads to a lower SINR estimate. Thus, to ensure the user-specific latency requirements (9b), the network strives to increase the users' rate over the available subset of RRUs $\mathcal{B}_k^c\ \forall k \in \mathcal{K}$, i.e., by consuming higher network power. Thus, it results in relatively lower average sum-power performance, as shown in Fig. 9. Conversely, a least pessimistic assumption on subset size (i.e., JT-CoMP, $L = B$) can provide higher instantaneous SINR (4), but it is more susceptible to the communication outage with the slight increase in blockage probability. Note that the problem (16) may also become infeasible for large values of link blockage probability due to the evolution of queue length buffer. More specifically, these outage events will eventually increase the queue-backlogs (31), i.e., due to unsuccessful data transmissions. Therefore, to guarantee the average latency requirements (9b), the network consumes much more power, and attempts to increase the achievable rates during potentially unblocked events. Thus, it results in a lower sum-power performance with increasing blockage, as shown in Fig. 9. Furthermore, our proposed dynamic choice of serving set size ensures the same average latency requirements with the minimum sum-power, and efficiently adopts with the uncertainties of mmWave radio channel and unpredictable blockage events. Note that the resulting gap in the dynamic subset selection around blockage probability $q = 20\%$ relative to the lower envelope, i.e., the parameter $L = 2$, is mainly due to the limited and discrete choices of parameter L, and also the approximation error of the time-averaged blockage in (29), as also illustrated in Fig. 7.

Fig. 10 illustrates the trade-offs between downlink rates and reliable mmWave connectivity with parameter $V = 1$ and blockage $q = 10\%$. However, similar behavior can be observed for different parameter settings, which is not included due to space limitations. It can concluded from Fig. 10 that the communication outage is decreased from 35% to less than 0.5% by changing parameter L from 4 (full JT-CoMP) to 2 (proposed), in problem (9). Thus, the pessimistic SINR estimate over the link blockage subset combinations of CoMP RRUs greatly improves the outage performance under the uncertainties of mmWave radio channel. Clearly, there is a trade-off between reliable connectivity, achievable rate, and sum-power performance. More specifically, for a given queue length (i.e., latency requirements), we can guarantee a user rate with minimum sum-power and vice-versa. However, compared to the baseline schemes, the proposed method significantly outperforms, and provides power-efficient and low-latency highly-reliable mmWave connectivity.

VII. CONCLUSION

In this paper, we have studied the trade-off between reliable downlink transmission and sum-power performance, in mmWave based time dynamic mobile networks, by exploiting the multi-antenna spatial diversity and CoMP connectivity. We considered an average sum-power minimization problem subject to maximum allowable queue length constraint per user. We have adapted the Lyapunov optimization framework, and derived a dynamic control algorithm for the long-term time-average stochastic problem. We proposed a robust transmit beamformer design by considering a pessimistic estimate of rates and a proactive selection of the serving set combinations of available CoMP RRUs. Furthermore, the non-convex and coupled constraints are handled using FP techniques. The closed-form algorithm is then provided by iteratively solving a system of KKT optimality conditions, while accounting for the uncertainties of mmWave radio channel. The numerical results manifested the robustness of the proposed beamformer design in the presence of random link blockages. Specifically, the achievable rate and sum-power performance with our proposed methods outperform the baseline scenarios while ensuring user-specific latency requirements, and thus, results in power-efficient and latency-aware highly-reliable mmWave communication.

ACKNOWLEDGMENT

This paper was presented in part at the IEEE Global Communications Conference (GLOBECOM), Taipei, Taiwan, in December 2020 [1].

REFERENCES

[1] D. Kumar, S. K. Joshi, and A. Tölli, "Latency-aware reliable mmWave communication via multi-point connectivity," in *Proc. IEEE Global Commun. Conf.*, 2020, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/9348026, doi: 10.1109/GLOBECOM42002.2020.9348026.

[2] N. Rajatheva *et al.*, "Scoring the Terabit/s goal: Broadband connectivity in 6G," 2020, *arXiv:2008.07220*.

[3] G. R. MacCartney, T. S. Rappaport, and S. Rangan, "Rapid fading due to human blockage in pedestrian crowds at 5G millimeter-wave frequencies," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2017, pp. 1–7.

[4] C. Slezak, V. Semkin, S. Andreev, Y. Koucheryavy, and S. Rangan, "Empirical effects of dynamic human-body blockage in 60 GHz communications," *IEEE Commun. Mag.*, vol. 56, no. 12, pp. 60–66, Dec. 2018.

[5] H. Shariatmadari, Z. Li, M. A. Uusitalo, S. Iraji, and R. Jäntti, "Link adaptation design for ultra-reliable communications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–5.

[6] D. Kumar, J. Kaleva, and A. Tolli, "Blockage-aware reliable mmWave access via coordinated multi-point connectivity," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4238–4252, Jul. 2021.

[7] J. Kaleva, A. Tolli, M. Juntti, R. A. Berry, and M. L. Honig, "Decentralized joint precoding with pilot-aided beamformer estimation," *IEEE Trans. Signal Process.*, vol. 66, no. 9, pp. 2330–2341, May 2018.

[8] *NR; Multi-Connectivity; Overall Description*, 3GPP, TSG RAN Meeting, document #82, TS 37.340, Jun. 15, 2018.

[9] Qualcomm. *How CoMP Can Extend 5G NR to High Capacity and Ultra-Reliable Communications*. Accessed: Apr. 22, 2021. [Online]. Available: https://www.qualcomm.com/documents/ how-comp-can-extend-5g-nr-high-capacity-and-ultra-reliable-communications

[10] A. Tolli, H. Pennanen, and P. Komulainen, "On the value of coherent and coordinated multi-cell transmission," in *Proc. IEEE Int. Conf. Commun. Workshops*, Jun. 2009, pp. 1–5.

[11] R. Irmer, H. Droste, P. Marsch, M. Grieger, G. Fettweis, S. Brueck, H.-P. Mayer, L. Thiele, and V. Jungnickel, "Coordinated multipoint: Concepts, performance, and field trial results," *IEEE Commun. Mag.*, vol. 49, no. 2, pp. 102–111, Feb. 2011.

[12] G. Nigam, P. Minero, and M. Haenggi, "Coordinated multipoint joint transmission in heterogeneous networks," *IEEE Trans. Commun.*, vol. 62, no. 11, pp. 4134–4146, Nov. 2014.

[13] G. R. MacCartney and T. S. Rappaport, "Millimeter-wave base station diversity for 5G coordinated multipoint (CoMP) applications," *IEEE Trans. Wireless Commun.*, vol. 18, no. 7, pp. 3395–3410, Jul. 2019.

[14] D. Maamari, N. Devroye, and D. Tuninetti, "Coverage in mmWave cellular networks with base station co-operation," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2981–2994, Apr. 2016.

[15] M. Gerasimenko, D. Moltchanov, M. Gapeyenko, S. Andreev, and Y. Koucheryavy, "Capacity of multiconnectivity mmWave systems with dynamic blockage and directional antennas," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3534–3549, Apr. 2019.

[16] M. Gapeyenko, V. Petrov, D. Moltchanov, M. R. Akdeniz, S. Andreev, N. Himayat, and Y. Koucheryavy, "On the degree of multi-connectivity in 5G millimeter-wave cellular urban deployments," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1973–1978, Feb. 2019.

[17] S. R. Khosravirad and H. Viswanathan, "Analysis of feedback error in automatic repeat reQuest," 2017, *arXiv:1710.00649*.

[18] Z. Ma, M. Xiao, Y. Xiao, Z. Pang, H. V. Poor, and B. Vucetic, "High-reliability and low-latency wireless communication for Internet of Things: Challenges, fundamentals, and enabling technologies," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 7946–7970, Oct. 2019.

[19] M. J. Neely, *Stochastic Network Optimization With Application to Communication and Queueing Systems* (Synthesis Lectures on Communication Networks), vol. 7. San Rafael, CA, USA: Morgan & Claypool, 2010.

[20] K. Shen and W. Yu, "Fractional programming for communication systems—Part I: Power control and beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, May 2018.

[21] T. Bai, R. Vaze, and R. W. Heath, Jr., "Analysis of blockage effects on urban cellular networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 9, pp. 5070–5083, Sep. 2014.

[22] M. Di Renzo, "Stochastic geometry modeling and analysis of multi-tier millimeter wave cellular networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 9, pp. 5038–5057, Sep. 2015.

[23] E. Hriba and M. C. Valenti, "The potential gains of macrodiversity in mmWave cellular networks with correlated blocking," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, Nov. 2019, pp. 171–176.

[24] A. Checko, H. L. Christiansen, Y. Yan, L. Scolari, G. Kardaras, M. S. Berger, and L. Dittmann, "Cloud RAN for mobile networks—A technology overview," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 1, pp. 405–426, 1st Quart., 2015.

[25] E. Nekouei, H. Inaltekin, and S. Dey, "Throughput scaling in cognitive multiple access with average power and interference constraints," *IEEE Trans. Signal Process.*, vol. 60, no. 2, pp. 927–946, Feb. 2012.

[26] D. Gross, J. F. Shortle, J. M. Thompson, and C. M. Harris, *Fundamentals of Queueing Theory*. Hoboken, NJ, USA: Wiley, 2008.

[27] K. Shen and W. Yu, "Fractional programming for communication systems—Part II: Uplink scheduling via matching," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2631–2644, May 2018.

[28] Q. Chen, K. Yang, H. Jiang, and M. Qiu, "Joint beamforming coordination and user selection for CoMP enabled NR-U networks," *IEEE Internet Things J.*, early access, Mar. 4, 2021.

[29] M. Grant and S. Boyd. (Mar. 2014). *CVX: Matlab Software for Disciplined Convex Programming, Version 2.1*. [Online]. Available: http://cvxr.com/cvx

[30] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[31] S. Boyd, L. Xiao, and A. Mutapcic, *Subgradient Methods* (lecture Notes of EE392o). Stanford, CA, USA: Stanford Univ., 2003.

[32] M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lebret, "Applications of second-order cone programming," *Linear Algebra Appl.*, vol. 284, nos. 1–3, pp. 193–228, Nov. 1998.

[33] D. Coppersmith and S. Winograd, "Matrix multiplication via arithmetic progressions," *J. Symbolic Comput.*, vol. 9, no. 3, pp. 251–280, Mar. 1990.

[34] G. Scutari, F. Facchinei, P. Song, D. P. Palomar, and J.-S. Pang, "Decomposition by partial linearization: Parallel optimization of multi-agent systems," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 641–656, Feb. 2014.

[35] A. M. Sayeed, "Deconstructing multiantenna fading channels," *IEEE Trans. Signal Process.*, vol. 50, no. 10, pp. 2563–2579, Oct. 2002.

DILEEP KUMAR (Graduate Student Member, IEEE) received the master's degree in communication engineering from the Indian Institute of Technology Bombay, India, in 2015. He is currently pursuing the Ph.D. degree with the University of Oulu, Finland. From 2015 to 2017, he worked at NEC Corporation, Tokyo, Japan, as a Research Engineer. In 2018, he joined the Centre for Wireless Communications (CWC), University of Oulu. His research interest includes signal processing for wireless communication systems.

SATYA KRISHNA JOSHI (Member, IEEE) received the M.Eng. degree in telecommunications from the Asian Institute of Technology, Pathumthani, Thailand, in 2007, and the Ph.D. degree from the University of Oulu, Finland, in 2018. His research interests include application of optimization techniques for signal processing and MIMO communications.

ANTTI TÖLLI (Senior Member, IEEE) received the Dr.Sc. (Tech.) degree in electrical engineering from the University of Oulu, Oulu, Finland, in 2008. From 1998 to 2003, he worked at Nokia Networks as a Research Engineer and a Project Manager both in Finland and Spain. In May 2014, he was granted a five year (2014–2019) Academy Research Fellow post by the Academy of Finland. During the academic year 2015–2016, he visited at EURECOM, Sophia Antipolis, France, while from August 2018 till June 2019, he was visiting at the University of California Santa Barbara, USA. He is currently Professor with the Centre for Wireless Communications (CWC), University of Oulu. He has authored numerous papers in peer-reviewed international journals and conferences and several patents all in the area of signal processing and wireless communications. His research interests include radio resource management and transceiver design for broadband wireless communications with a special emphasis on distributed interference management in heterogeneous wireless networks. From 2017 to 2021, he served as an Associate Editor for IEEE Transactions on Signal Processing.