Joint Load Balancing and Interference Mitigation in 5G Heterogeneous Networks

Trung Kien Vu, *Student Member, IEEE,* Mehdi Bennis, *Senior Member, IEEE,*
Sumudu Samarakoon, *Student Member, IEEE,* Mérouane Debbah, *Fellow, IEEE,*
and Matti Latva-aho, *Senior Member, IEEE*

***Abstract*—We study the problem of joint load balancing and interference mitigation in heterogeneous networks (HetNets) in which massive multiple-input multiple-output (MIMO) macro cell base station (BS) equipped with a large number of antennas, overlaid with wireless self-backhauled small cells (SCs) are assumed. Self-backhauled SC BSs with full-duplex communication employing regular antenna arrays serve both macro users and SC users by using the wireless backhaul from macro BS in the same frequency band. We formulate the joint load balancing and interference mitigation problem as a network utility maximization subject to wireless backhaul constraints. Subsequently, leveraging the framework of stochastic optimization, the problem is decoupled into dynamic scheduling of macro cell users, backhaul provisioning of SCs, and offloading macro cell users to SCs as a function of interference and backhaul links. Via numerical results, we show the performance gains of our proposed framework under the impact of SCs density, number of BS antennas, and transmit power levels at low and high frequency bands. It is shown that our proposed approach achieves a $5.6\times$ gain in terms of cell-edge performance as compared to the closed-access baseline in ultra-dense networks with 350 SC BSs per km^2.**

***Index Terms*—Massive MIMO, ultra dense small cells, mmWave communications, self-backhaul, full-duplex, imperfect CSI, random matrix theory, non-convex optimization.**

I. Introduction

To meet the massive data traffic demands in next generation 5G wireless networks a number of emerging technologies are currently investigated: 1) higher frequency spectrum (mmWave); 2) advanced spectral-efficiency techniques (massive MIMO); and 3) ultra-dense small cell deployments [2]. In this paper, we focus on the interplay between massive MIMO and a dense deployment of SCs in higher frequency bands. Massive MIMO plays an important role in wireless networks due to an improvement in energy and spectral efficiency [3]. In massive MIMO, a macro base station (MBS) equipped with a few hundreds antennas simultaneously serves tens of user equipments (UEs) and provides wireless backhaul to SCs, while the remaining degree of freedom of massive MIMO can be used to mitigate the cross-tier interference. Ultra dense SC deployment provides an effective solution to increase network capacity by a factor of $100\times$ or more and offloads the wireless data from the MBS [4]. In order to reduce the deployment cost of SC, wireless backhaul has been considered as an attractive solution. In parallel to that, recent advances in full-duplex (FD) enables doubling spectral efficiency and lowering latency in which FD-enabled SCs relay data from the massive MIMO MBS to the UEs in the same frequency band [5].

MmWave with short wavelength enables Massive MIMO to pack more antennas into highly directional footprint and to smartly do beamforming [6], making Massive MIMO practically feasible in real deployments. Recently, the efficiency of combining massive MIMO and in-band wireless backhaul-based SC networks was studied in [5], [7], focusing on minimizing power consumption. The problem of user association for load balancing in heterogeneous networks (HetNets) has been studied in [8]. Although, users can be associated to more than one BS in order to reduce the load on the macro cell, deploying ultra-dense small cell networks makes user association more challenging. The work in [8] did not consider other important aspects in 5G such as Massive MIMO, FD-enabled SCs, and mmWave communications. Recent work in [9] has addressed the user-cell association for Massive MIMO HetNets, which did not consider the joint optimization of load balancing, precoder design, and power allocation. Also the wireless backhaul faces the problem of limited-backhaul; hence the backhaul constraint needs to be considered. Thus far, the key challenge of how to dynamically optimize the overall network performance taking into account the backhaul dynamics and constraints, and load balancing utilizing the combination of Massive MIMO, FD-enabled SCs, and mmWave communications has not been fully addressed [10].

User association taking into account dynamic backhaul in 5G HetNets faces a new challenge due to self-backhauled SCs, i.e., guaranteeing wireless backhaul capacity between MBS

Manuscript received Nov 14, 2016; revised April 09, 2017 and May 26, 2017; accepted June 13, 2017; Date of publication June 27, 2017; The authors would like to thank the Finnish Funding Agency for Technology and Innovation (Tekes), Nokia, Huawei, and Anite for project funding. The Academy of Finland funding through the grant 284704 and the Academy of Finland CARMA project are also acknowledged. The research of M. Debbah has been supported by the ERC Starting Grant 305123 MORE (Advanced Mathematical Tools for Complex Network Engineering). This paper was presented in part at the 22th European Wireless Conference, Oulu, Finland, May 2016 [1]. The associate editor coordinating the review of this paper and approving it for publication was S. Jin (*Corresponding author: Trung Kien Vu.)*

T. K. Vu, S. Sumudu, and M. Latva-aho are with the Centre for Wireless Communications, University of Oulu, 90014 Oulu, Finland, (email: {trungkien.vu, sumudu.samarakoon, 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 (LANEAS), CentraleSupélec, Université Paris-Saclay, 91192 Gif-sur-Yvette, France and is with the Mathematical and Algorithmic Sciences Laboratory, Huawei France R&D, 92100 Paris, France (e-mail: merouane.debbah@huawei.com).

Citation information: DOI: 10.1109/TWC.2017.2718504, IEEE Transaction on Wireless Communications

and SCs in order to offload the traffic from MBSs to SCs. It raises the following important question: Should MBS serve all macro UEs (MUEs) even though it is highly loaded or offload some MUEs to SCs subject to the wireless backhaul capacity? Due to the random deployment of massive number of devices, UEs around hotspots (i.e. airport lounges, shopping malls, stations, and other crowded places) may receive poor services from a-far-MBS with multiple beams focused on the same location. On the contrary, these UEs will receive better services from nearby SCs with a reliable wireless backhaul composed of strong single beam from the MBS or multiple received antennas at SCs.

A. Main Contributions

The main contributions of this work are to study the problem of joint load balancing, interference mitigation, and in-band wireless backhauling taking into account dynamic backhaul and traffic load, which are listed as follows:

- The problem of joint load balancing (user association and user scheduling) and interference management (beamforming design and power allocation) for 5G HetNets is modeled in which a DL scheduler is designed at the MBS to schedule macro UEs and provide backhaul to in-band FD-enabled SCs, with FD capability SCs serve both MUEs and small cell UEs in the same frequency band. Moreover, an interference management scheme is proposed to mitigate both co-tier and cross-tier interference from the MBS and FD-enabled SCs by designing a hierarchical precoding scheme and controlling the transmission of SCs. The problem is cast as a network utility maximization (NUM) problem subject to dynamic wireless backhaul constraints, traffic load, and imperfect channel state information (CSI). To make problem tractable, by invoking results from random matrix theory (RMT), we derive a closed-form expression of the signal-to-interference-plus-noise-ratio (SINR) and transmit power when the numbers of MBS antennas and users grow very large.
- A Lyapunov framework is applied in order to solve the NUM problem in polynomial time. The NUM problem is decomposed into dynamic scheduling of MUEs, backhaul provisioning of FD-enabled SCs, and offloading MUEs to FD-enabled SCs. The joint load balancing and operation mode (FD or half-duplex) subproblem, which is a non-convex program with binary variables, is converted into a convex program by using the successive convex approximation (SCA) method. The motivations of using SCA are due to (i) its low complexity and fast convergence, and (ii) the obtained solution which yields many relaxed variables is close to zero or one.
- A performance evaluation is carried out to compare the proposed algorithm with other baselines under the impact of SCs density, number of BS antennas, and transmit power levels at low/high frequency bands. The effect of pilot training and channel aging is also studied to show the performance of Massive MIMO.
- A comprehensive performance analysis of our proposed algorithm based on the Lyapunov framework is provided. There exists an $[O(1/v), O(v)]$ utility-queue backlog tradeoff, which leads to an utility-delay balancing [11], where v is the Lyapunov control parameter. Moreover, a convergence analysis of the approximation method based on the SCA method is studied.

B. Related Work

The authors in [12] addressed the problem of dynamic resource control for HetNets with flexible backhaul (wired and wireless). However, the problem of load balancing when the number of antennas and users grows large is not considered. The user association problem has been studied for HetNets in [8], [9], which does not take into account backhaul constraints. As pointed out in [10], [13] the current solutions for user association problem ignore the backhaul constraints, which is very crucial since the capacity of open access SCs with either wired or wireless backhaul always faces the limited backhaul constraint. Moreover, the load balancing problem should take into account imperfect CSI due to mobility, which is ignored in the previous work. Our previous work in [1] has considered the problem of joint in-band scheduling and interference mitigation in 5G HetNets without considering the user association. In this work, we extend [1] by considering the load balancing problem taking into account the backhaul constraint and imperfect CSI, and further provides insights into the performance analysis of our proposed algorithm based on the Lyapunov framework and convergence of the SCA method.

The rest of this paper is organized as follows.[1] Section II describes the system model and Section III provides the problem formulation for load balancing and interference mitigation. Section IV introduces the Lyapunov framework used to solve our problem. In Section V, we present the numerical results. We conclude the paper in Section VI.

II. System Model

A. System Model

The downlink (DL) transmission of a HetNet scenario is considered as shown in Fig. 1 in which a MBS b_0 is underlaid with a set of uniformly deployed S FD-enabled SCs, $\mathcal{S} = \{b_s | s \in \{1, \ldots, S\}\}$. Let $\mathcal{B} = \{b_0\} \cup \mathcal{S}$ denote the set of all base stations, where $|\mathcal{B}| = 1 + S$. The MBS is equipped with N number of antennas and serves a set of single-antenna M MUEs $\mathcal{M} = \{1, \ldots, M\}$. Let $\mathcal{K} = \mathcal{M} \cup \mathcal{S}$ denote the set of users associated with MBS b_0, where $|\mathcal{K}| = K = M + S$. The user indices $k = 1, 2, ..., M$ represent the corresponding MUEs indices $m = 1, 2, ..., M$, while the user indices $k = M+1, M+2, ..., M+S$ represent the corresponding SCs indices $s = 1, 2, ..., S$. We assume open access policy at FD-enabled SCs and each FD-enabled SC is equipped with N_s+1 antennas: one receiving antenna is used for the wireless backhaul and

[1]The lowercase letters, boldface lowercase letters, (boldface) uppercase letters and italic boldface uppercase letters are used to represent scalars, vectors, matrices, and sets, respectively. $\mathbf{X}^\dagger$ and rank($\mathbf{X}$) denote the Hermitian transpose and the rank of matrix $\mathbf{X}$, respectively. diag($x_1, x_2, ...x_N$) denotes the block diagonal matric whose diagonal blocks are given by $x_1, x_2, ..., x_N$ and the identity matrix of size N is denoted by $\mathbf{I}_N$. The cardinality of a set $\mathcal{S}$, is denoted by $|\mathcal{S}|$. $\mathcal{CN}(0, \sigma^2)$ denotes the Gaussian random distribution of zero mean and variance of σ^2.

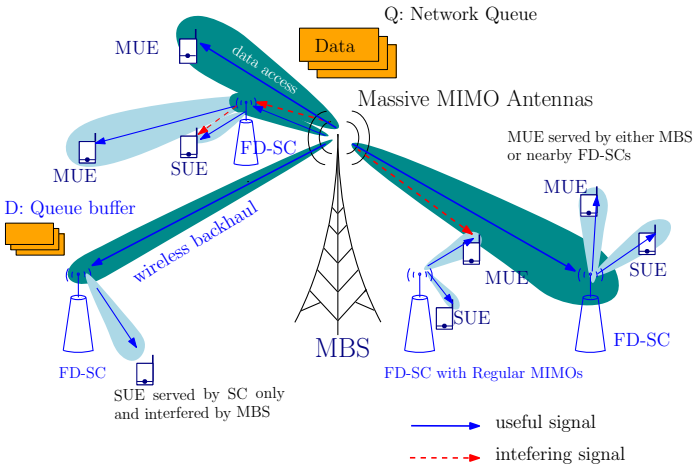

Fig. 1. Integrated access and backhaul architecture for the considered 5G network scenario.

N_s transmitting antenna to serve its single-antenna small cell UEs (SUEs) or other MUEs at the same frequency band. Let $C = \{c_1, c_2, \ldots, c_S\}$ denote the set of SUEs, where $|C| = S$. Moreover, SCs are assumed to be FD capable with perfect self-interference cancelation (SIC) capabilities[2]. Co-channel time-division duplexing (TDD) protocol is considered in which the MBS and FD-enabled SCs share the entire bandwidth, and the DL transmission occurs at the same time. In this work, we consider a large number of antennas at both macro and SC BSs and a dense deployment of MUEs and SCs, such that $M, N, N_s, S \gg 1$.

B. Channel Model

We denote $\mathbf{h}_m^{(b_0)} = \left[h_m^{(b_0,1)}, h_m^{(b_0,2)}, \cdots, h_m^{(b_0,N)}\right]^T \in \mathbb{C}^{N\times 1}$ the propagation channel between the mth MUE and the antennas of the MBS b_0 in which $h_m^{(b_0,n)}$ is the channel between the mth MUE and the nth MBS antenna. Let $\mathbf{H}^{(b_0),\mathrm{M}} = \left[\mathbf{h}_1^{(b_0)}, \mathbf{h}_2^{(b_0)}, \cdots, \mathbf{h}_M^{(b_0)}\right] \in \mathbb{C}^{N\times M}$ denote the channel matrix between all MUEs and the MBS antennas. Moreover, we assume imperfect CSI for MUEs due to mobility and we denote $\hat{\mathbf{H}}^{(b_0),\mathrm{M}} = \left[\hat{\mathbf{h}}_1^{(b_0)}, \hat{\mathbf{h}}_2^{(b_0)}, \cdots, \hat{\mathbf{h}}_M^{(b_0)}\right] \in \mathbb{C}^{N\times M}$ as the estimate of $\mathbf{H}^{(b_0),\mathrm{M}}$ in which the imperfect CSI can be modeled as [14]:

$$\hat{\mathbf{h}}_m^{(b_0)} = \sqrt{N\boldsymbol{\Theta}_m^{(b_0)}}\,\hat{\mathbf{w}}_m^{(b_0)}, \tag{1}$$

where $\hat{\mathbf{w}}_m^{(b_0)} = \sqrt{1-{\tau_m}^2}\mathbf{w}_m^{(b_0)} + \tau_m \mathbf{z}_m^{(b_0)}$ is the estimate of the small-scale fading channel matrix and $\boldsymbol{\Theta}_m^{(b_0)}$ is the spatial channel correlation matrix that accounts for path loss and shadow fading. Here, $\mathbf{w}_m^{(b_0)}$ and $\mathbf{z}_m^{(b_0)}$ are the real channel and the channel noise, respectively, modeled as Gaussian random matrix with zero mean and variance $1/N$. The channel estimate error of MUE m is denoted by $\tau_m, \tau_m \in [0,1]$; in case of perfect CSI, $\tau_m = 0$. Similarly, let $\mathbf{H}^{(b_0),\mathrm{S}} \in \mathbb{C}^{N\times S}$ and $\mathbf{H}^{(b_0),\mathrm{C}} \in \mathbb{C}^{N\times S}$ denote the channel matrices from the MBS antennas to SCs and SUEs, respectively. Let $\mathbf{h}_u^{(b_s)} \in \mathbb{C}^{N_s\times 1}$ denote the channel propagation from SC b_s to any receiver u. Let c_s denote the SUE served by the SC b_s.

[2]The case of imperfect SIC is left for future work.

III. Load Balancing and Interference Mitigation

In this section, we formulate the joint optimization of user association, user scheduling, beamforming design, and power allocation. To that end, we first derive the received signal, data rate, and power transmit for each receiver (SCs are also treated as macro BS's UEs). We then formulate the problem as a network utility maximization subject to wireless backhaul constraints. However, the formulated problem does not have closed-form expressions for the objective and constraints. Hence, we apply RMT [15] to get these closed-form expressions. We finally utilize the tool of stochastic optimization to decouple our problem into several solvable sub-problems.

The problem of user scheduling and user association for load balancing in the DL is addressed in which the MBS simultaneously provides data transmission to MUEs and wireless backhaul to the FD-enabled SCs, while the SCs with FD capability serve both SUEs and MUEs. For each MUE $m \in \mathcal{M}$, let binary variable $l_m^{(b_s)}$ indicate the transmission association from BS $b_s \in \mathcal{B}$ to MUE m, i.e., $l_m^{(b_s)} = 1$ when MUE m is associated with BS b_s, otherwise $l_m^{(b_s)} = 0$. Similarly, let binary variables $l_{s+M}^{(b_0)}$ and $l_{c_s}^{(b_s)}$ denote the transmission association indicators from MBS b_0 to SC s and from SC b_s to SUE c_s, respectively. We assume that each MUE m connects to one BS (either MBS b_0 or SC b_s) at time slot t. Each SC is equipped with N_s transmitting antennas, and we assume that each SC serves up to N_s^{au} active users (either SUE or MUE) at each time slot, such that $N_s^{\mathrm{au}} \le N_s$, where the superscript au stands for "active users". Hence, we have the following constraints for load balancing:

$$\textstyle\sum_{s=0}^{S} l_m^{(b_s)} \le 1, \sum_{m=1}^{M} l_m^{(b_s)} + l_{c_s}^{(b_s)} \le N_s^{\mathrm{au}}, \forall\ s, m \in \mathcal{K}. \tag{2}$$

We define vector $\mathbf{l} = \left\{l_j^{(b_s)} | b_s \in \mathcal{B}, j \in \{\mathcal{M} \cup \mathcal{S} \cup C\}\right\}$ containing all transmission indicators between BSs and UEs. Let $N_s^{\mathrm{tx}} = \sum_{m=1}^{M} l_m^{(b_s)} + l_{c_s}^{(b_s)}$ be the total number of transmissions at SC, where superscript tx stands for "transmissions", and thus the latter of (2) becomes $N_s^{\mathrm{tx}} \le N_s^{\mathrm{au}}, \forall s \in \mathcal{S}$.

A. Downlink Transmission Signal

The MBS serves two types of users: MUEs with imperfect CSI and FD-enabled SCs with perfect CSI. Let $p_m^{(b_0)}$, $p_{s+M}^{(b_0)}$, and $P^{(b_0)}$ denote the DL MBS transmit power assigned to MUE m, the DL MBS transmit power assigned to SC s, and the maximum transmit power at the MBS, respectively. We focus on the multiple-input single-output (MISO) channel, where the MBS with N antennas can serve K UEs. Here, we take into account user scheduling and association, and our proposal can apply to any special case when number of UEs is larger than number of antennas, i.e., $K > N$. SC exploits FD capability to double capacity, FD-enabled SC causes unwanted FD interference: cross-tier interference to adjacent MUEs (or other SCs), and co-tier interference to other UEs. Hence, in order to convert the interference channel to the MISO channel, we design a precoder at the MBS and propose an operation mode policy to control FD interference in order to treat the total FD interference as additional noise.

Definition 1: [Operation Mode Policy] We define $\boldsymbol{\beta}$ as the operation mode to control the FD-enabled SC transmission to

reduce FD interference. The operation mode is expressed as $\boldsymbol{\beta}(t) = \{\beta^{(b_s)}(t) \mid \beta^{(b_s)}(t) \in \{0,1\}, \forall s \in \mathcal{S}\}$. Here, $\beta^{(b_s)}(t) = 1$ indicates SC b_s operates in FD mode and $\beta^{(b_s)}(t) = 0$ for half-duplex (HD) mode.

We assume that the MBS uses a precoding scheme, $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_K] \in \mathbb{C}^{N\times K}$. To exploit the degrees of freedom of massive MIMO, the hierarchical interference mitigation scheme in [16], [17] is applied to design the precoder, i.e., $\mathbf{V} = \mathbf{UT}$, where $\mathbf{T} \in \mathbb{C}^{N\times N_{\text{itf}}}$ is used to control co-tier interference and capture the spatial multiplexing gain, and $\mathbf{U} \in \mathbb{C}^{N_{\text{itf}}\times K}$ is used to suppress cross-tier interference. Here, $N_{\text{itf}} < N$, where the subscript itf stands for "interference". The precoder $\mathbf{U}$ is chosen such that

$$\mathbf{U}^\dagger \textstyle\sum_{s=1}^{S} \beta^{(b_s)} \boldsymbol{\Theta}_s^{(b_0)} = 0, \tag{3}$$

where $\boldsymbol{\Theta}_s^{(b_0)} \in \mathbb{C}^{N\times N}$ is the sum of the correlation matrices between MBS antennas and users belong to SC s. Here, $\mathbf{U}$ is in the null space of $\sum_{s=1}^{S} \beta^{(b_s)} \boldsymbol{\Theta}_s^{(b_0)}$. Note that $\beta^{(b_s)}$ determines that the transmission of FD-enabled SC is enabled or not. The precoder $\mathbf{T}$ is designed to adapt to the real time CSI based on $\hat{\mathbf{H}}^\dagger\mathbf{U} \in \mathbb{C}^{K\times N_{\text{itf}}}$, where $\hat{\mathbf{H}} = [\hat{\mathbf{h}}^{(b_0)}]_{k\in\mathcal{K}}^\dagger$. In this paper, we consider the regularized zero-forcing (RZF) precoding[3] that is given by $\mathbf{T} = \left(\mathbf{U}^\dagger\hat{\mathbf{H}}^\dagger\hat{\mathbf{H}}\mathbf{U} + N\alpha\mathbf{I}_{N_{\text{itf}}}\right)^{-1}\mathbf{U}^\dagger\hat{\mathbf{H}}^\dagger$, where the regularization parameter $\alpha > 0$ is scaled by N to ensure that the matrix $\mathbf{U}^\dagger\hat{\mathbf{H}}^\dagger\hat{\mathbf{H}}\mathbf{U} + N\alpha\mathbf{I}_{N_{\text{itf}}}$ is well conditioned as $N \to \infty$. The precoder $\mathbf{T}$ is chosen to satisfy the power constraint $\text{Tr}\left(\mathbf{PT}^\dagger\mathbf{T}\right) \le P^{(b_0)}$, where $\mathbf{P} = \text{diag}(p_1^{(b_0)}, p_2^{(b_0)}, \ldots, p_K^{(b_0)})$. We also assume that each SC uses ZF precoding to server its users, $\mathbf{F}^{(b_s)} = [\mathbf{f}_1^{(b_s)}, \mathbf{f}_2^{(b_s)}, \ldots, \mathbf{f}_{N_s^{\text{tx}}}^{(b_s)}] \in \mathbb{C}^{N_s\times N_s^{\text{tx}}}$ which reads $\mathbf{f}_u^{(b_s)} = \mathbf{h}_u^{(b_s)\dagger}\left(\mathbf{h}_u^{(b_s)}\mathbf{h}_u^{(b_s)\dagger}\right)^{-1}$ such that $\mathbf{F}^{(b_s)}$ is chosen to satisfy the equality power constraint $\text{Tr}\left(\mathbf{P}^{(b_s)}\mathbf{F}^{(b_s)\dagger}\mathbf{F}^{(b_s)}\right) = P^{(b_s)}$[4]. Here, $\mathbf{P}^{(b_s)} = \text{diag}(p_1^{(b_s)}, p_2^{(b_s)}, \ldots, p_{N_s^{\text{tx}}}^{(b_s)})$. The channel propagation from the SC b_s to the MUE m (referred to as user u) is $\mathbf{h}_u^{(b_s)} = \hat{\mathbf{h}}_m^{(b_s)} = \sqrt{N_s\boldsymbol{\Theta}_m^{(b_s)}}\left(\sqrt{1-{\tau_m}^2}\mathbf{w}_m^{(b_s)} + \tau_m\mathbf{z}_m^{(b_s)}\right)$, where $\boldsymbol{\Theta}_m^{(b_s)} \in \mathbb{C}^{N_s\times N_s}$ is the channel correlation matrix. Here, $\mathbf{w}_m^{(b_s)}$ and $\mathbf{z}_m^{(b_s)}$ are the real channel and the channel noise from SC b_s to MUE m, respectively, modeled as a Gaussian random matrix with zero mean and variance $1/N_s$.

By utilizing a massive number of antennas at MBS, a large spatial degree of freedom is utilized to serve MUEs and FD-enabled SCs, while the remaining degrees of freedom are used to mitigate cross-tier interference. In massive MIMO system, the total number of antennas is considered as the degree of freedom [16]. Hence, we have the antenna constraint for user association and operation mode such that $\sum_{k=1}^{K} l_k^{(b_0)}(t) + \sum_{s=1}^{S} N_s^{\text{tx}}(t) \le N$. For notational simplicity, we remove the time dependency from the symbols throughout the discussion. The received signal $y_m^{(b_0)}$ at each MUE $m \in \mathcal{M}$ at time instant t is given by

$$\begin{aligned} y_m^{(b_0)} = {} & l_m^{(b_0)}\sqrt{p_m^{(b_0)}}\mathbf{h}_m^{(b_0)\dagger}\mathbf{v}_m x_m^{(b_0)} \\ & + \underbrace{\sum_{s=1}^{S}\beta^{(b_s)}\textstyle\sum_{u=1}^{N_s^{\text{tx}}} l_u^{(b_s)}\sqrt{p_u^{(b_s)}}\mathbf{h}_m^{(b_s)\dagger}\mathbf{f}_u^{(b_s)}x_u^{(b_s)}}_{\text{cross-tier interference}} \\ & + \underbrace{\sum_{k=1,k\neq m}^{K} l_k^{(b_0)}\sqrt{p_k^{(b_0)}}\mathbf{h}_m^{(b_0)\dagger}\mathbf{v}_k x_k^{(b_0)}}_{\text{co-tier interference}} + \eta_m, \end{aligned} \tag{4}$$

where $x_m^{(b_0)}$ is the signal symbol from the MBS to the MUE m, $\mathbf{v}_m$ is the precoding vectors of MUE m, and $\eta_m \sim \mathcal{CN}(0,1)$ is the thermal noise at MUE m. While $x_u^{(b_s)}$ is the transmit signal symbol from SC b_s to its user u.

At time instant t, the received signal $y_{s+M}^{(b_0)}$ at each SC $s \in \mathcal{K}$ suffers from self-interference, cross-tier and co-tier interference, which is given by

$$\begin{aligned} y_{s+M}^{(b_0)} = {} & l_{s+M}^{(b_0)}\sqrt{p_{s+M}^{(b_0)}}\mathbf{h}_{s+M}^{(b_0)\dagger}\mathbf{v}_{s+M}x_{s+M}^{(b_0)} \\ & + \underbrace{\sum_{s'=1,s'\neq s}^{S}\beta^{(b_{s'})}\sum_{u'=1}^{N_{s'}^{\text{tx}}} l_{u'}^{(b_{s'})}\sqrt{p_{u'}^{(b_{s'})}}\mathbf{h}_s^{(b_{s'})\dagger}\mathbf{f}_{u'}^{(b_{s'})}x_{u'}^{(b_{s'})}}_{\text{cross-tier interference}} \\ & + \underbrace{\beta^{(b_s)}\sum_{u=1}^{N_s^{\text{tx}}} l_u^{(b_s)}\sqrt{p_u^{(b_s)}}\mathbf{h}_s^{(b_s)\dagger}\mathbf{f}_u^{(b_s)}x_u^{(b_s)}}_{\text{self-interference}} \\ & + \underbrace{\sum_{k=1,k\neq s+M}^{K} l_k^{(b_0)}\sqrt{p_k^{(b_0)}}\mathbf{h}_{s+M}^{(b_0)\dagger}\mathbf{v}_k x_k^{(b_0)}}_{\text{co-tier interference}} + \eta_{s+M}, \end{aligned} \tag{5}$$

where $x_{s+M}^{(b_0)}$ is the signal symbol from the MBS to the SC s, $\mathbf{v}_{s+M}$ is the precoding vectors of SC s, and $\eta_{s+M} \sim \mathcal{CN}(0,1)$ is the thermal noise of the SC s.

The received signal from the SC b_s at receiver u, $y_u^{(b_s)} = 0$, if the SC b_s operates in HD mode, $\beta^{(b_s)} = 0$. For FD mode, $\beta^{(b_s)} = 1$, the received signal $y_u^{(b_s)}$ is given by

$$\begin{aligned} y_u^{(b_s)} = {} & \beta^{(b_s)}l_u^{(b_s)}\sqrt{p_u^{(b_s)}}\mathbf{h}_u^{(b_s)\dagger}\mathbf{f}_u^{(b_s)}x_u^{(b_s)} \\ & + \underbrace{\sum_{s'=1,s'\neq s}^{S}\beta^{(b_{s'})}\sum_{u'=1,}^{N_{s'}^{\text{tx}}} l_{u'}^{(b_{s'})}\sqrt{p_{u'}^{(b_{s'})}}\mathbf{h}_u^{(b_{s'})\dagger}\mathbf{f}_{u'}^{(b'_s)}x_{u'}^{(b_{s'})}}_{\text{co-tier interference}} \\ & + \underbrace{\beta^{(b_s)}\sum_{j=1,j\neq u}^{N_s^{\text{tx}}} l_j^{(b_s)}\sqrt{p_j^{(b_s)}}\mathbf{h}_u^{(b_s)\dagger}\mathbf{f}_j^{(b_s)}x_u^{(b_s)}}_{\text{co-tier self-interference}} \\ & + \underbrace{\sum_{k=1,k\neq u}^{K} l_k^{(b_0)}\sqrt{p_k^{(b_0)}}\mathbf{h}_u^{(b_0)\dagger}\mathbf{v}_k x_k^{(b_0)}}_{\text{cross-tier interference}} + \eta_u, \end{aligned} \tag{6}$$

where $x_u^{(b_s)}$ is the transmit data symbol from the SC b_s to receiver u and $\eta_u \sim \mathcal{CN}(0,1)$ is the thermal noise at receiver

[3]Other precoders are left for future work.

[4]We choose the equality constraints for transmit power at SCs to reach the optimal rate at maximum power rather than using $\text{Tr}\left(\mathbf{P}^{(b_s)}\mathbf{F}^{(b_s)\dagger}\mathbf{F}^{(b_s)}\right) \le P^{(b_s)}$, since the power at SCs is relatively small.

u. We imply that the receiver u can be either a SUE or an MUE.

The precoder $\mathbf{V}$ is designed at the MBS to null the co-tier interference and to remove completely the cross-tier interference to SCs's users (3) and the self-interference is well treated, while $\mathrm{Tr}\big(\mathbf{P}^{(b_s)}\mathbf{F}^{(b_s)\dagger}\mathbf{F}^{(b_s)}\big) = P^{(b_s)}$. Thus, according to (4)-(6), the SINRs of an MUE m served by MBS, a SC s served by MBS, a receiver u served by SC are given in (7)-(9), respectively.

B. Joint Load Balancing and Interference Mitigation Algorithm

Let us consider a joint optimization of load balancing $\mathbf{l}$, operation mode $\boldsymbol{\beta}$, interference mitigation $\mathbf{U}$, and transmit power allocation $\mathbf{p} = (p_1^{(b_0)}, p_2^{(b_0)}, \ldots, p_K^{(b_0)})$ that satisfies the transmit power budget of MBS i.e. , $\mathrm{Tr}\big(\mathbf{P}\mathbf{T}^\dagger\mathbf{T}\big) \leq P^{(b_0)}$. We define $\zeta_k^{(b_s)} = \frac{P^{(b_s)}|\mathbf{h}_k^{(b_s)\dagger}|^2}{|\eta_k|^2}$ and ϵ_o as the FD interference to noise ratio (INR) from FD-enabled SC b_s to any scheduled receiver k, and the allowed FD INR threshold, respectively. The FD interference threshold is defined such that $\sum_{k=1}^{K}\sum_{s=1}^{S} l_k^{(b_0)}\beta^{(b_s)}\zeta_k^{(b_s)} \leq \epsilon_o$, such that the total FD interference is considered as noise. Under the operation mode policy, we schedule the receiver i and enable the transmission of SC b_s as long as $\sum_{k=1}^{K}\sum_{s=1}^{S} l_k^{(b_0)}\beta^{(b_s)}\zeta_k^{(b_s)} \leq \epsilon_o$. Let $\mathbf{\Lambda}^o = \{\mathbf{l}, \boldsymbol{\beta}\}$ be a composite control variable of user association and operation mode. We define $\mathbf{\Lambda} = \{\mathbf{\Lambda}^o, \mathbf{U}, \mathbf{p}\}$ as a composite control variable, which adapts to the spatial channel correlation matrix $\mathbf{\Theta}$.

For a given $\mathbf{\Lambda}$ that satisfies (3) and operation mode policy, the respective Ergodic data rates of SC s and SUE u are $r_{s+M}(\mathbf{\Lambda}|\mathbf{\Theta}) = \mathbb{E}\big[\log\big(1+\gamma_{s+M}^{(b_0)}\big)\big]$ and $r_u^{(b_s)}(\mathbf{\Lambda}|\mathbf{\Theta}) = \mathbb{E}\big[\log\big(1+\gamma_u^{(b_s)}\big)\big]$. While from the constraint (2) the Ergodic data rate of MUE m will depend on which BS the MUE is associated with, i.e., $r_m(\mathbf{\Lambda}|\mathbf{\Theta}) = \mathbb{E}\big[\log\big(1+\gamma_m^{(b_0)}\big)\big] + \sum_{s=1}^{S}\min\{\mathbb{E}\big[\log\big(1+\gamma_m^{(b_s)}\big)\big], \quad r_s(\mathbf{\Lambda}|\mathbf{\Theta}) - \sum_{u\neq m} r_u^{(b_s)}(\mathbf{\Lambda}|\mathbf{\Theta})\}$. In other words, the first term is the data rate from from the MBS to MUE when MUE is associated with the MBS, while the second term is when the FD-enabled SCs allow MUE to connect (If MUE is connected to the FD-enabled SC, then the rate of MUE should be the minimum between $r_m^{(b_s)}(\mathbf{\Lambda}|\mathbf{\Theta})$ and data stream from the MBS via FD-enabled SC to MUE, excepts other SC's users).

Definition 2: For any vector $\mathbf{x}(t) = (x_1(t), ..., x_K(t))$, let $\bar{\mathbf{x}} = (\bar{x}_1, \cdots, \bar{x}_K)$ denote the time average expectation of $\mathbf{x}(t)$, where $\bar{\mathbf{x}} \triangleq \lim_{t\to\infty}\frac{1}{t}\sum_{\tau=0}^{t-1}\mathbb{E}[\mathbf{x}(\tau)]$. Similarly, $\bar{\mathbf{r}} \triangleq \lim_{t\to\infty}\frac{1}{t}\sum_{\tau=0}^{t-1}\mathbb{E}[\mathbf{r}(\tau)]$ denotes the time average expectation of the Ergodic data rate.

For a given composite control variable $\mathbf{\Lambda}$ that adapts to the spatial channel correlation matrix $\mathbf{\Theta}$, the average data rate region is defined as the convex hull of the average data rate of users, which is expressed as:

$$\begin{aligned}\mathcal{R} \triangleq \Big\{&\bar{\mathbf{r}}(\mathbf{\Lambda}|\mathbf{\Theta}) \in \mathbf{R}_+^K \mid \mathbf{l} \in \{0,1\}^{K+MS+S}, \boldsymbol{\beta} \in \{0,1\}^S,\\ &\textstyle\sum_{s=0}^{S} l_m^{(b_s)} \leq 1, \quad \forall\ m \in \mathcal{M},\\ &\textstyle\sum_{m=1}^{M} l_m^{(b_s)} + l_{c_s}^{(b_s)} = N_s^{\mathrm{tx}}, N_s^{\mathrm{tx}} \leq N_s^{\mathrm{au}}, \quad \forall\ b_s \in \mathcal{S},\\ &\textstyle\sum_{k=1}^{K} l_k^{(b_0)} + \sum_{s=1}^{S} N_s^{\mathrm{tx}} \leq N,\\ &\textstyle\sum_{k=1}^{K}\sum_{s=1}^{S} l_k^{(b_0)}\beta^{(b_s)}\zeta_k^{(b_s)} \leq \epsilon_o,\\ &\mathrm{Tr}\big(\mathbf{P}\mathbf{T}^\dagger\mathbf{T}\big) \leq P^{(b_0)},\ \mathbf{U}^\dagger\textstyle\sum_{s=1}^{S}\beta^{(b_s)}\mathbf{\Theta}_s^{(b_0)} = 0\Big\},\end{aligned}$$

where $\bar{\mathbf{r}}(\mathbf{\Lambda}|\mathbf{\Theta}) = (\bar{r}_1(\mathbf{\Lambda}|\mathbf{\Theta}), \ldots, \bar{r}_K(\mathbf{\Lambda}|\mathbf{\Theta}))^T$. Following the results from [18], the boundary points of the rate regime with total power constraint and no self-interference are Pareto-optimal[5]. Moreover, according to [19, Proposition 1], if the INR covariance matrices approach the identity matrix, the Pareto rate regime of the MIMO interference system is convex. Hence, our rate regime is a Pareto-optimal, and thus is convex with above constraints.

Let us assume that each FD-enabled SC acts as a relay to forward data to its users. If the MBS transmits data to FD-enabled SC b_s, but the transmission of SC b_s is disabled, it cannot serve its SUE. Hence, we define $\mathbf{D}(\mathbf{t}) = (D_1(t), D_2(t), \ldots, D_S(t))$ as a data queue at SCs, where at each time slot t, the wireless backhaul queue at FD-enabled SC b_s is

$$D_s(t+1) = \max[D_s(t) + r_{s+M}(t) - r_{c_s}^{(b_s)}(t), \quad 0], \quad \forall\ s \in \mathcal{S}. \tag{10}$$

The SC offloads some MUEs from the MBS if the wireless backhaul capacity between the SCs and the MBS is guaranteed, and hence, for each SC we have the following wireless backhaul condition for all $t \geq 0$: "If the access link between the MUE m and the MBS is better than the link between the MUE m and the SCs, then the MUE connects with the MBS rather than with other SCs", i.e.,[6]

$$\text{if} \quad r_{s+M}(t) \leq r_m^{(b_0)}(t), \quad \text{then} \quad l_m^{(b_s)} = 0, \quad \forall s \in \mathcal{S}, m \in K. \tag{11}$$

Definition 3: [Queue stability] For any discrete queue $Q(t)$ over time slots $t \in \{0, 1, \ldots\}$ and $Q(t) \in R_+$, $Q(t)$ is stable if $\bar{Q} \triangleq \lim_{t\to\infty}\frac{1}{t}\sum_{\tau=0}^{t-1}\mathbb{E}\big[|Q(\tau)|\big] < \infty$. A queue network is stable if each queue is stable.

We define the network utility function $f_0(\cdot)$ to be non-decreasing, concave over the convex region $\mathcal{R}$ for a given $\mathbf{\Theta}$. The objective is to maximize the network utility under wireless backhaul constraints and imperfect CSI. Thus, the NUM problem is given by,

$$\max_{\bar{\mathbf{r}}} \quad f_0(\bar{\mathbf{r}}) \tag{12a}$$

$$\text{subject to} \quad (11), \quad \bar{\mathbf{r}} \in \mathcal{R}, \quad \bar{\mathbf{D}} < \infty, \tag{12b}$$

[5]The Pareto optimal is the set of user rates at which it is impossible to improve any of the rates without simultaneously decreasing at least one of the others.

[6]The queues of MUEs are handled at the MBS and SCs strictly handle data for SUEs, hence when SCs open connection for MUEs, they should have immediate capacity in terms of data rate. We do not include the constraint (11) for the closed access case in [1].

$$\gamma_m^{(b_0)} = \frac{l_m^{(b_0)} p_m^{(b_0)} |\mathbf{h}_m^{(b_0)\dagger} \mathbf{v}_m|^2}{\sum_{k\neq m} l_k^{(b_0)} p_k^{(b_0)} |\mathbf{h}_m^{(b_0)\dagger} \mathbf{v}_k|^2 + \sum_s \beta^{(b_s)} P^{(b_s)} |\mathbf{h}_m^{(b_s)\dagger}|^2 + 1}. \tag{7}$$

$$\gamma_{s+M}^{(b_0)} = \frac{l_{s+M}^{(b_0)} p_{s+M}^{(b_0)} |\mathbf{h}_{s+M}^{(b_0)\dagger} \mathbf{v}_{s+M}|^2}{\sum_{k\neq s+M} l_k^{(b_0)} p_k^{(b_0)} |\mathbf{h}_{s+M}^{(b_0)\dagger} \mathbf{v}_k|^2 + \sum_{s'\neq s} \beta^{(b_{s'})} P^{(b_{s'})} |\mathbf{h}_s^{(b_{s'})\dagger}|^2 + 1}. \tag{8}$$

$$\gamma_u^{(b_s)} = \frac{\beta^{(b_s)} l_u^{(b_s)} p_u^{(b_s)} |\mathbf{h}_u^{(b_s)\dagger} \mathbf{f}_u^{(b_s)}|^2}{\beta^{(b_s)} \sum_{j=1, j\neq u} l_j^{(b_s)} p_j^{(b_s)} |\mathbf{h}_u^{(b_s)\dagger} \mathbf{f}_j^{(b_s)}|^2 + \sum_{s'\neq s} \beta^{(b_{s'})} P^{(b_{s'})} |\mathbf{h}_u^{(b_{s'})\dagger}|^2 + 1}. \tag{9}$$

where $f_0(\bar{\mathbf{r}}) = \sum_{k=1}^K \omega_k(t) f(\bar{r}_k)$ with $\omega_k(t) \geq 0$ is the weight of user k, $f(\cdot)$ is assumed to be twice differentiable, concave, and increasing L-Lipschitz function for all $\bar{r} \geq 0$. Solving (12) is non-trivial since the average rate region $\mathcal{R}$ does not have a tractable form. To overcome this challenge, we need to find closed-form expressions of the data rate and the average transmit power. Inspired by [15], we invoke RMT to get the closed-form expressions for the user data rate and transmit power as $N \gg K$.

C. Closed-Form Expression via Deterministic Equivalent

We invoke recent results from RMT in order to get the deterministic equivalent of user rate and transmit power via **Theorem 1**.

Theorem 1: Recall that α is the RZF parameter. As $N \gg K$; $N, K \to \infty$, by applying the technique in [15, Theorem 2], the deterministic equivalent of the asymptotic SINR of MUE m is

$$\gamma_m^{(b_0)} \xrightarrow{a.s.} \frac{l_m^{(b_0)} p_m^{(b_0)} (1-\tau_m^2)(\Omega_m)^2}{\Phi},$$

where $\xrightarrow{a.s.}$ denotes the almost sure convergence and $\Phi = \Upsilon_m \left[\alpha^2 - \tau_m^2 \left(\alpha^2 - (\alpha+\Omega_m)^2\right)\right] + (\alpha+\Omega_m)^2 (1 + \sum_{s=1}^S \beta^{(b_s)} \zeta_m^{(b_s)})$. Here, $\Omega_m = \frac{1}{N} \mathrm{Tr}(\tilde{\mathbf{\Theta}}_m \mathbf{G})$ forms the unique positive solution of which is the Stieltjes transform of nonnegative finite measure [15, Theorem 1], where $\mathbf{G} = \left(\frac{1}{N} \sum_{k=1}^K \frac{\tilde{\mathbf{\Theta}}_k}{\alpha+\Omega_k} + \mathbf{I}_{N_{\mathrm{itf}}}\right)^{-1}$. In addition, $\Upsilon_m = \frac{1}{N} \sum_{k=1, k\neq m}^K \frac{\alpha^2 l_k^{(b_0)} p_k^{(b_0)} e_{km}}{(\alpha+\Omega_k)^2}$, and $\tilde{\mathbf{\Theta}}_k = \mathbf{U}\mathbf{U}^\dagger \mathbf{\Theta}_k^{(b_0)} \mathbf{U}\mathbf{U}^\dagger$. $\mathbf{e} = [e_k], k \in \mathcal{K}$, and $\mathbf{e_m} = [e_{mk}], k \in \mathcal{K}$ are given by $\mathbf{e} = (\mathbf{I} - \mathbf{J})^{-1}\mathbf{u}$, $\mathbf{e_k} = (\mathbf{I} - \mathbf{J})^{-1}\mathbf{u_k}$, where $\mathbf{J} = [J_{ij}], i, j \in \mathcal{K}$. $\mathbf{u} = [u_k], k \in \mathcal{K}$, $\mathbf{u_m} = [u_{mk}], k \in \mathcal{K}$ are given by $\mathbf{J}_{ij} = \frac{\frac{1}{N}\mathrm{tr}\tilde{\mathbf{\Theta}}_i \mathbf{G} \tilde{\mathbf{\Theta}}_j \mathbf{G}}{N(\alpha+\Omega_j)^2}$, $u_{mk} = \frac{1}{\alpha^2 N} \mathrm{tr}\tilde{\mathbf{\Theta}}_k \mathbf{G} \tilde{\mathbf{\Theta}}_m \mathbf{G}$, $u_k = \frac{1}{\alpha^2 N} \mathrm{tr}\tilde{\mathbf{\Theta}}_k \mathbf{G}^2$. Similarly, the SINR of SC b_s is

$$\gamma_s^{(b_0)} \xrightarrow{a.s.} \frac{l_s^{(b_0)} p_s^{(b_0)} (\Omega_s)^2}{\alpha^2 \Upsilon_s + (\alpha+\Omega_s)^2 (1 + \sum_{s'=1, s'\neq s}^S \beta^{(b_{s'})} \zeta_s^{(b_{s'})})}.$$

The power constraint at the MBS can be calculated as $\frac{1}{N} \sum_{k=1}^K \frac{p_k^{(b_0)} \alpha^2 e_k}{(\alpha+\Omega_k)^2} - P^{(b_0)} \leq 0$. Moreover, following the analysis in the proof of [15, Theorem 3], [16, Lemma 6] for a small fixed[7] $\alpha > 0$, $\Upsilon_k = \boldsymbol{O}(1)$ and $\alpha^2 e_k = \Omega_k + \boldsymbol{O}(\alpha)$ yield the deterministic equivalent of the asymptotic SINRs of UEs (7)-(9) as

$$\gamma_m^{(b_0)}(\mathbf{\Lambda}|\mathbf{\Theta}) \xrightarrow{a.s.} \frac{l_m^{(b_0)} p_m^{(b_0)} (1-\tau_m^2)}{1 + \sum_{s=1}^S \beta^{(b_s)} \zeta_m^{(b_s)}}, \tag{13}$$

$$\gamma_s^{(b_0)}(\mathbf{\Lambda}|\mathbf{\Theta}) \xrightarrow{a.s.} \frac{l_s^{(b_0)} p_s^{(b_0)}}{1 + \sum_{s'=1, s'\neq s}^S \beta^{(b_{s'})} \zeta_s^{(b_{s'})}}, \tag{14}$$

$$\gamma_u^{(b_s)}(\mathbf{\Lambda}|\mathbf{\Theta}) \xrightarrow{a.s.} \frac{\beta^{(b_s)} l_u^{(b_s)} p_u^{(b_s)}}{1 + \sum_{s'=1, s'\neq s}^S \beta^{(b_{s'})} \zeta_u^{(b_{s'})}}. \tag{15}$$

Moreover, we obtain the closed-form expression for the transmit power constraint, i.e.,

$$\frac{1}{N} \sum_{k=1}^K \frac{p_k^{(b_0)}}{\Omega_k} - P^{(b_0)} \leq 0.$$

Although the closed-form expressions of average data rate and transmit power are obtained, our problem considers a time-average optimization with a large number of control variables, and dynamic traffic load over the convex region for a given composite control variable $\mathbf{\Lambda}$ and $\mathbf{\Theta}$. Our aim is to maximize the aggregate network utility subject to queue stability in which the well-known Lyapunov optimization yields an utility throughput optimality and stability [20]. Hence, we apply the drift-plus-penalty technique [20] to solve load balancing, operation mode selection, and power allocation problems.

IV. Lyapunov Optimization Framework

The network operation is modeled as a queueing network that operates in discrete time $t \in \{0, 1, 2, \dots\}$. Let $a_k(t)$ denote the bursty data arrival destined for each user k, i.i.d over time slot t. Let $\mathbf{Q}(t)$ denote the vector of transmission queue blacklogs at MBS at slot t. The queue evolution is given by

$$Q_k(t+1) = \max\left[Q_k(t) - r_k(t), \quad 0\right] + a_k(t), \quad \forall\ k \in \mathcal{K}. \tag{16}$$

Here, we consider the bound of the traffic arrival of user k is bounded such that $0 \leq a_k(t) \leq a_k^{\max}$, for some constant $a_k^{\max} < \infty$. Futuremore, let $r_k^{\max}(t)$ be the upper bound of data rate for user k at time slot t, such that $r_k^{\max}(t) \leq a_k^{\max}$. The set at constraint (12b) is replaced by an another equivalent set by introducing auxiliary variables $\boldsymbol{\varphi}(t) \in \mathcal{R}$, $\boldsymbol{\varphi}(t) = \left(\varphi_1(t), \dots, \varphi_K(t)\right)$ that satisfies $\bar{\varphi}_k \leq \bar{r}_k$, where $\bar{\varphi}_k \triangleq \lim_{t\to\infty} \frac{1}{t} \sum_{\tau=0}^{t-1} \mathbb{E}\left[\varphi_k(\tau)\right]$. The evolution of wireless backhaul queue is rewritten as

$$D_s(t+1) = \max\left[D_s(t) + \varphi_{s+M}(t) - r_{c_s}^{(b_s)}(t), \quad 0\right], \quad \forall\ s \in \mathcal{S}. \tag{17}$$

[7]The deterministic equivalent holds for a small fixed α as studied in [16], while the problem of finding the optimal value α has been studied in [15], [17].

For a given $\mathbf{\Lambda}$ and $\mathbf{\Theta}$, the optimization problem (12) subject to the network stability and dynamic backhaul can be posed as

$$\min_{\bar{\boldsymbol{\varphi}}} \quad -f_0(\bar{\boldsymbol{\varphi}}) \tag{18a}$$

$$\text{subject to} \quad \bar{\varphi}_k - \bar{r}_k \leq 0, \quad \forall\ k \in \mathcal{K}, \tag{18b}$$

$$(11), \quad \bar{\mathbf{D}} < \infty, \bar{\mathbf{Q}} < \infty. \tag{18c}$$

In order to ensure the inequality constraint (18b), we introduce a virtual queue vector $Y(t)$ which evolves as follows

$$Y_k(t+1) = \max\ [Y_k(t) + \varphi_k(t) - r_k(t), \quad 0], \quad \forall\ k \in \mathcal{K}. \tag{19}$$

We define the queue backlog vector as $\Sigma(t) = \big[\mathbf{Q}(t), \mathbf{Y}(t), \mathbf{D}(t)\big]$ (whereas the stability of $\Sigma(t)$ yields all constraints of problem (18) are hold). The Lyapunov function can be written as

$$L(\Sigma(t)) \triangleq \frac{1}{2}\big[\ \textstyle\sum_{k=1}^{K} Q_k(t)^2 + \sum_{k=1}^{K} Y_k(t)^2 + \sum_{s=1}^{S} D_s(t)^2\big].$$

For each time slot t, $\mathbf{\Delta}(\Sigma(t))$ denotes the Lyapunov drift, which is given by

$$\mathbf{\Delta}(\Sigma(t)) \triangleq \mathbb{E}\big[L(\Sigma(t+1)) - L(\Sigma(t))|\Sigma(t)\big].$$

Noting that $\max[a,0]^2 \leq a^2$ and $(a \pm b)^2 \leq a^2 \pm 2ab + b^2$ for any real positive number a, b, and thus, by neglecting the index t we have:

$$(\max\ [Q_k - r_k, \quad 0] + a_k)^2 - Q_k^2 \leq 2Q_k(a_k - r_k) + (a_k - r_k)^2,$$

$$\max\ [Y_k + \varphi_k - r_k, \quad 0]^2 - Y_k^2 \leq 2Y_k(\varphi_k - r_k) + (\varphi_k - r_k)^2,$$

$$\max\ [D_s + \varphi_{s+M} - r_{c_s}^{(b_s)}(t), \quad 0]^2 - D_s^2 \leq 2D_s(\varphi_{s+M} - r_{c_s}^{(b_s)}(t)) + (\varphi_{s+M} - r_{c_s}^{(b_s)}(t))^2.$$

We assume that $\boldsymbol{\varphi}_k \in \mathcal{R}$ and a feasible $\mathbf{l}$ for all t and all possible $\Sigma(t)$, we have

$$\begin{aligned}\mathbf{\Delta}(\Sigma(t)) \leq\ & \Psi + \textstyle\sum_{k=1}^{K} Q_k(t)\mathbb{E}\Big[a_k(t) - r_k(t)|\Sigma(t)\Big] \\ & + \textstyle\sum_{s=1}^{S} D_s(t)\mathbb{E}\big[\varphi_{s+M}(t) - r_{c_s}^{(b_s)}(t)|\Sigma(t)\big] \\ & + \textstyle\sum_{k=1}^{K} Y_k(t)\mathbb{E}\big[\varphi_k(t) - r_k(t)|\Sigma(t)\big].\end{aligned} \tag{20}$$

Here $\mathbf{\Delta}(\Sigma(t)) \leq \Pi$, where Π represents the R.H.S of (20), and Ψ is a finite constant that satisfies $\Psi \geq \frac{1}{2}\sum_{k=1}^{K}\mathbb{E}\big[\big(a_k(t) - r_k(t)\big)^2|\Sigma(t)\big] + \frac{1}{2}\sum_{k=1}^{K}\mathbb{E}\big[\big(\varphi_k(t) - r_k(t)\big)^2|\Sigma(t)\big] + \frac{1}{2}\sum_{s=1}^{S}\mathbb{E}\big[\big(\varphi_{s+M}(t) - r_s^{c_s}(t)\big)^2|\Sigma(t)\big]$, for all t and all possible $\Sigma(t)$. We apply the Lyapunov drift-plus-penalty technique [20], where the solution of (18) is obtained by minimizing the Lyapunov drift and a penalty from the objective function, i.e.,

$$\min\ \Pi - \nu\mathbb{E}[f_0(\boldsymbol{\varphi}(t))].$$

Here, the parameter ν is chosen as non-negative constant to control optimal minimization solution [20]. Since Ψ is finite, the problem is to minimize the below expression subject to the convex set hull, given by (21). Note that (21) is decoupled over user association, user scheduling, and operation mode variables (2⋆), auxiliary variables (3⋆), and precoder and power allocation variables (1⋆), respectively as in (21). Hence, the respective variables can be found independently by minimizing the individual term at each time. Fig. 2 summarizes the relationship among various subproblems.

A. Joint Load Balancing and Operation Mode Selection

First, the problem of joint load balancing and FD-enabled SC operation mode selection in (2⋆) is cast as the minimization problem below.

$$\begin{aligned}\min_{\mathbf{l},\boldsymbol{\beta}}\ & -\textstyle\sum_{k=1}^{K} A_k(t)\log\big(1 + l_k^{(b_0)}(t)\frac{p_k^{(b_0)}(1-\tau_k^2)}{1+\sum_{s=1}^{S}\beta^{(b_s)}\zeta_k^{(b_s)}}\big) \\ & - \textstyle\sum_{s=1}^{S} D_s(t)\log\big(1 + \beta^{(b_s)}(t)\frac{l_{c_s}^{(b_s)}(t)p_{c_s}^{(b_s)}}{1+\sum_{s'\neq s}^{S}\beta^{(b_{s'})}\zeta_{c_{s'}}^{(b_{s'})}}\big)\end{aligned} \tag{22a}$$

$$\text{subject to} \quad l_j^{(b_s)}(t) \in \{0,1\}, \forall\ j \in \mathcal{K}\cup\mathcal{C}, \quad \forall\ b_s \in \mathcal{B}, \tag{22b}$$

$$\beta^{(b_s)}(t) \in \{0,1\}, N_s^{\text{tx}}(t) \leq N_s^{\text{au}}, \forall\ s \in \mathcal{S}, \tag{22c}$$

$$\textstyle\sum_{s=0}^{S} l_m^{(b_s)}(t) \leq 1, \forall\ m \in \mathcal{M},$$

$$\textstyle\sum_{m=1}^{M} l_m^{(b_s)}(t) + l_{c_s}^{(b_s)}(t) = N_s^{\text{tx}}(t), \tag{22d}$$

$$(11), r_k(t) \in \mathcal{R},$$

$$\textstyle\sum_{k=1}^{K} l_k^{(b_0)}(t) + \sum_{s=1}^{S} N_s^{\text{tx}}(t) \leq N, \tag{22e}$$

$$\textstyle\sum_{k=1}^{K}\sum_{s=1}^{S} l_k^{(b_0)}(t)\beta^{(b_s)}(t)\zeta_k^{(b_s)}(t) \leq \epsilon_o, \tag{22f}$$

where $A_k(t) = Q_k(t) + Y_k(t)$. This problem is a non-convex program with binary variables. It turns out this problem has a hidden convexity structure and the non-convex terms can be iteratively approximated by its convex upper bound via an iterative SCA method. The motivations of utilizing the SCA method are due to (i) its low complexity and fast convergence [21, Lemma 3.5] and (ii) the obtained solution which yields many relaxed variables are close to zero or one [22]. In this regard, we convexify this problem to find a sub optimal solution. First, we relax the binary constraints (22b) and (22c) to linear constraints as continuous variables. Secondly, at each iteration i the non-convex constraint (22f) is approximated by upper convex approximation, i.e.,

$$\sum_{k=1}^{K}\sum_{s=1}^{S}\Big(\frac{\lambda_{ks}^{(i)}(l_k^{(b_0)}(t))^2}{2} + \frac{(\beta^{(b_s)})^2(t)}{2\lambda_{ks}^{(i)}}\Big)\zeta_k^{(b_s)}(t) - \epsilon_o \leq 0,$$

for every fixed positive value $\lambda_{ks}^{(i)}$. Finally, instead of minimizing the non-convex objective function (22a) we convert it into a convex function by the followings. We minimize its upper bound by replacing the denominators, i.e., $1 + \sum_{s=1}^{S}\beta^{(b_s)}\zeta_m^{(b_s)}$ with largest bound, i.e., $1 + \epsilon_0$. Due to interference constraint (22f), we obtain the upper bound as below

$$\begin{aligned}& -\sum_{k=1}^{K} A_k(t)\log\big(1 + \frac{l_k^{(b_0)}(t)p_k^{(b_0)}(1-\tau_k^2)}{1+\epsilon_0}\big) \\ & -\sum_{s=1}^{S} D_s(t)\log\big(1 + \beta^{(b_s)}(t)\frac{l_{c_s}^{(b_s)}(t)p_{c_s}^{(b_s)}}{1+\epsilon_0}\big).\end{aligned}$$

Using the similar approach as convexifying the interference constraint (22f), we convexify the second part of these objective function which still remains non-convex. We denote the lower bound of SINR of UE served SC b_s as $\underline{\gamma}^{b_s}(t)$, let us set $\tilde{l}_{c_s}^{(b_s)}(t) \triangleq \frac{l_{c_s}^{(b_s)}(t)p_u^{(b_s)}}{1+\epsilon_0}$. Then we have:

$$\underline{\gamma}^{b_s}(t) \leq \beta^{(b_s)}(t)\tilde{l}_{c_s}^{(b_s)}(t), \quad \forall\ s \in \mathcal{S}, \tag{23}$$

$$\Big[\big[\overbrace{-\textstyle\sum_k \big(Q_k(t)+Y_k(t)\big)r_k(\mathbf{\Lambda}(t))}^{\text{Impact of network queue, virtual queue, and } \mathbf{\Lambda}}\big]_{1\star}\overbrace{-\textstyle\sum_s D_s(t)r_{c_s}^{(b_s)}(\boldsymbol{\beta}^{(b_s)}(t))}^{\text{Impact of SC queue and } \boldsymbol{\beta}}\Big]_{2\star}+\Big[\overbrace{\textstyle\sum_k Y_k(t)\varphi_k(t)+\sum_s D_s(t)\varphi_{s+M}(t)}^{\text{Impact of virtual queue, SC queue, and auxiliaries}}\ \overbrace{-\nu f_0(\boldsymbol{\varphi}(t))}^{\text{penalty}}\Big]_{3\star}. \tag{21}$$

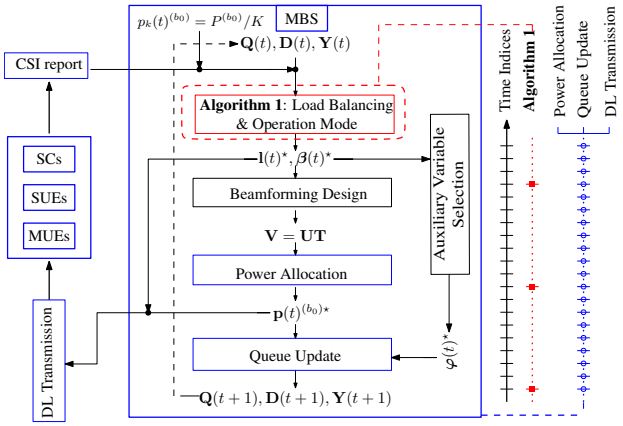

Fig. 2. Joint load balancing and interference mitigation algorithm.

by introducing the new slack variable $\iota_s^2(t)$, (23) is equivalent to:

$$\frac{1}{4}\big(\beta^{(b_s)}(t)-\tilde{l}_{c_s}^{(b_s)}(t)\big)^2+\iota_s^2(t)\le\frac{1}{4}\big(\beta^{(b_s)}(t)+\tilde{l}_{c_s}^{(b_s)}(t)\big)^2, \tag{24}$$

$$\text{and}\quad \underline{\gamma}^{b_s}(t)\le\iota_s^2(t),\quad \forall\ s\in\mathcal{S}. \tag{25}$$

where the constraint (24) holds a form of the second-order cone inequalities (SOC), while the RHS of the set of constraints in (25) are still non-convex, which can be approximated by using the iterative SCA method [21]. We rewrite the constraint (25) as

$$\underline{\gamma}^{b_s}(t)\le\hat{\iota}_s^{(i)2}(t)+2\hat{\iota}_s^{(i)}(t)(\iota_s(t)-\hat{\iota}_s^{(i)}(t)),\quad \forall\ s\in\mathcal{S}, \tag{26}$$

where at iteration $i+1$, we update $\hat{\iota}_s^{(i+1)}(t)$ such that $\hat{\iota}_s^{(i+1)}(t)=\iota_s^{(i)}(t)$. Hence, the optimal value of $\mathbf{\Lambda}^o$ is given by

$$\begin{aligned}
\min_{\mathbf{l},\boldsymbol{\beta}}\quad & -\textstyle\sum_{k=1}^K A_k(t)\log\big(1+l_k^{(b_0)}(t)\frac{p_k^{(b_0)}(1-\tau_k^2)}{1+\epsilon_0}\big)\\
& -\textstyle\sum_{s=1}^S D_s(t)\log\big(1+\underline{\gamma}^{b_s}(t)\big) && \text{(27a)}\\
\text{subject to}\quad & l_j^{(b_s)}(t)\in[0,1],\forall\ j\in\mathcal{K}\cup C,\forall\ b_s\in\mathcal{B}, && \text{(27b)}\\
& \beta^{(b_s)}(t)\in[0,1],N_s^{\text{tx}}(t)\le N_s^{\text{au}},\forall\ s\in\mathcal{S}, && \text{(27c)}\\
& (22d),(22e),(24),(26), && \text{(27d)}\\
& \textstyle\sum_{k=1}^K\sum_{s=1}^S\big(\frac{\lambda_{ks}^{(i)}(l_k^{(b_0)}(t))^2}{2}\\
& +\frac{(\beta^{(b_s)})^2(t)}{2\lambda_{ks}^{(i)}}\big)\zeta_k^{(b_s)}(t)-\epsilon_o\le 0. && \text{(27e)}
\end{aligned}$$

At each time slot t, the approximated problem (27) is iteratively solved as in Algorithm 1. We numerically observe that the SCA-based Algorithm 1 converges quickly within few iterations and yields a continuous relaxation solution of many user association and operation mode variables close or equal to binary. To ensure that all users will be served, when performing Algorithm 1 each user is assumed to receive the same transmit power to find the best scheduled users. Moreover, the scheduling will be performed in a long-term period, while the power allocation problem is executed in a short-term period. Since the objective function of the problem (27)

Algorithm 1 Joint load balancing and operation mode algorithm

Initialization $i:=0$, $\lambda_{ks}^{(i)},\hat{\iota}_s^{(i)}:=$ randomly positive that satisfy all constraints.
repeat
Solve (27) with $\lambda_{ks}^{(i)},\hat{\iota}_s^{(i)}$ to get optimal value $\mathbf{\Lambda}^{o\star}=\{\mathbf{l}^\star,\boldsymbol{\beta}^\star\}$.
Update $\mathbf{\Lambda}^{o(i)}:=\mathbf{\Lambda}^{o\star}$ and $\lambda_{ks}^{(i+1)}:=\frac{\beta^{(b_s)(i)}}{l_k^{(b_0)(i)}}$; $\hat{\iota}_s^{(i+1)}:=\iota_s^{(i)}$; $i:=i+1$.
until Convergence

is a maximum weighted matching problem with respect to linear or square function, we use a low-complexity binary search algorithm [23] to obtain the final solutions with lower dimensions. Let $K_1=\{j,s|l_j^{(b_s)\star},\beta^{(b_s)\star}=1\}$, $K_{\text{uct}}=\{j,s|\xi\le l_j^{(b_s)\star},\beta^{(b_s)\star}\le 1\}$, and $K_0=\{j,s|l_j^{(b_s)\star},\beta^{(b_s)\star}\le\xi\}$ denote set of selected variables, set of uncertain variables, set of removed variables, respectively, where ξ is some small threshold. First, we determine the set K_1, K_{uct}, and K_0 based on ξ. Then, we consider to select among the uncertain variables in K_{uct}. By sorting K_{uct} in a descending order, a loop starts by selecting one by one variable based on their largest weights according to the objective function. We set the value uncertain variable to 1, and add it to K_1, if it satisfies the antennas (27d) and interference (27e) constraints. If it does not satisfy the constrains, we add it to K_0. The loop stops until reaching the last uncertain variable or the antennas constrain is over. Finally, K_1 is kept, while K_0 and K_{uct} are removed.

B. The Selection of Auxiliary Variable

The optimal auxiliary variable from (3⋆) is computed by

$$\min_{\boldsymbol{\varphi}(t)}\textstyle\sum_{k=1}^K Y_k(t)\varphi_k(t)+\sum_{s=1}^S D_s(t)\varphi_{s+M}(t) \tag{28a}$$

$$-\nu\textstyle\sum_{k=1}^K\omega_k(t)f(\varphi_k(t)) \tag{28b}$$

$$\text{subject to}\quad \varphi_k(t)\le a_k^{\max}(t). \tag{28c}$$

Since the above optimization problem is convex, let $\varphi_k^*(t)$ be the optimal solution obtained by the first order derivative of the objective function of (28). With a logarithmic utility function, we have:

$$\varphi_k^*(t)=\begin{cases}\frac{\nu\omega_k(t)}{Y_k(t)} & \text{if } k\le M,\\ \frac{\nu\omega_k(t)}{Y_k(t)+D_{k-M}(t)} & \text{otherwise.}\end{cases}$$

The optimal auxiliary variable is $\min\{\varphi_k^*(t),a_k^{\max}(t)\}$.

C. Interference Mitigation and Power Allocation

For given scheduled users, the precoder $\mathbf{U}$ is found by solving (3). Finally, problem (18) is decomposed to find the transmit power $p_k^{(b_0)}(t)$ from (1⋆) that minimizes:

$$\min_{\mathbf{p}(t)} \quad -\sum_{k=1}^{K} A_k(t) r_k(\mathbf{p}(t)) \tag{29a}$$

$$\text{subject to} \quad \frac{1}{N}\sum_{k=1}^{K} \frac{p_k^{(b_0)}(t)}{\Omega_k(t)} - P^{(b_0)} \leq 0,$$

$$p_k^{(b_0)}(t) \geq 0, \forall\ k \in \mathcal{K}.$$

The objective function (29) is rewritten as $n(\mathbf{p}(t)) = -\sum_{k=1}^{K} A_k(t) \log\left(1 + p_k^{(b_0)}(t) n_k(t)\right)$, where $n_k(t) = \frac{l_k^{(b_0)}(t)(1-\tau_k^2)}{1+\sum_{s=1}^{S} \beta^{(b_s)}(t)\zeta_k^{(b_s)}(t)}$. The objective function is strictly convex for $p_k^{(b_0)}(t) \geq 0, \forall k \in \mathcal{K}$, and the constraints are compact. Hence, the optimal solution of $\mathbf{p}^{\star}(t)$ exists, the Lagrangian function is written as $\mathcal{L}(\mathbf{p}(t), \mu_0) = n(\mathbf{p}(t)) + \mu_0 \mathbf{g}(\mathbf{p}(t))$, where $\mu_0 \geq 0$ is the KKT multiplier. The KKT conditions are

$$\nabla n(\mathbf{p}(t))^T + \mu_0 \frac{1}{N}\sum_{k=1}^{K} \frac{1}{\Omega_k(t)} = 0. \tag{30}$$

$$\mu_0 \left(\frac{1}{N}\sum_{k=1}^{K} \frac{p_k^{(b_0)}(t)}{\Omega_k(t)} - P^{(b_0)} \right) = 0. \tag{31}$$

$$\frac{1}{N}\sum_{k=1}^{K} \frac{p_k^{(b_0)}(t)}{\Omega_k(t)} - P^{(b_0)} \leq 0, \qquad -\mathbf{p}(t) \leq 0, \mu_0 \geq 0. \tag{32}$$

Here, $\nabla n(\mathbf{p}(t))^T = (n'(p_1^{(b_0)}(t)), \ldots, n'(p_K^{(b_0)}(t)))$ where $n'(p_k^{(b_0)}(t)) = \frac{-A_k(t) n_k(t)}{1+p_k^{(b_0)}(t) n_k(t)}$. For $\mu_0 \neq 0$, from (30), obtaining

$$p_k^{(b_0)}(t) = \max[\frac{A_k N \Omega_k(t)}{\mu_0} - \frac{1}{n_k(t)}, 0], \tag{33}$$

from (31) and (33) we derive μ_0. Finally, the optimal value of $p_k(t)^{(b_0)\star}$ is obtained with (33).

D. Queue Update

Update the virtual queues $Y_k(t)$ and $D_s(t)$ according to (19) and (17), and the actual queue $Q_k(t)$ in (16).

Theorem 3 is provided to show the performance analysis of network utility maximization based on Lyapunov framework and prove that the queues are stable.

Theorem 2: [Optimality] Assume that all queues are initially empty. For arbitrary arrival rates, the operation mode and load balancing is chosen to satisfy (21) and the rate regime. For a given constant $\chi \geq 0$, the network utility maximization with any $\nu > 0$ provides the following utility performance with $\chi - approximation$

$$f_0 \geq f_0^* - \frac{\Psi + \chi}{\nu},$$

where $f_0^{\star}$ is the optimal network utility over the rate regime. While the strong stability of the virtual queues and the network queues is given by

$$Q_k(t) \leq \nu\omega_k(t)\pi_k + 2a_k^{\max}, \quad \forall t \geq 0, \quad \forall k \in \mathcal{K},$$

$$Y_k(t) \leq \nu\omega_k(t)\pi_k + a_k^{\max}, \quad \forall t \geq 0, \quad \forall k \in \mathcal{K},$$

$$D_s(t) \leq \nu\omega_{s+M}(t)\pi_{s+M} + a_{s+M}^{\max}, \quad \forall t \geq 0, \quad \forall s \in \mathcal{K}.$$

Proof: Proof can be found in [20] and is omitted for the sake of brevity.

E. Relaxation of Utility Function

Note that the previous discussion explains how to transform the above non-convex program (22) as a generic convex program. Although it can be solved by using the modern solvers, generally it requires more computation time. In order to reduce the computation time and speed up the optimization convergence, we relax the log function of the objective function (27a) by a set of linear functions. Moreover, in order to model and solve the problem efficiently, we use YALMIP toolbox [24], which can employ SDPT3 [25] or MOSEK [26] as internal solver. In general, we rewrite the log function as

$$1 + \gamma(l_k) \geq e^{r_k},$$

where $\gamma(l_k)$ is the SINR as a function of l_k. By using the results of approximation of second order cone programming [27], [28], (34) can be approximated by a set of following linear equations

$$\begin{aligned}
&1 + \gamma(l_k) \geq \kappa_0, \quad 1 + \kappa_1 \geq \|[1 - \kappa_1 \quad 2 + r_k/2^{i-1}]\|_2,\\
&1 + \kappa_2 \geq \|[1 - \kappa_2 \quad 5/3 + r_k/2^{i}]\|_2,\\
&1 + \kappa_3 \geq \|[1 - \kappa_3 \quad 2\kappa_1]\|_2,\\
&\kappa_4 \geq \kappa_2 + \kappa_3/24 + 19/72,\\
&1 + \kappa_j \geq \|[1 - \kappa_j \quad 2\kappa_{j-1}]\|_2, \quad j = \{5, \cdots, i+3\},\\
&1 + \kappa_0 \geq \|[1 - \kappa_0 \quad 2\kappa_{i+3}]\|_2,
\end{aligned}$$

where $\{\kappa_j\}_{j=0,1,\cdots,i+3}$, are new introduced variables, and the accuracy of the approximation depends on i. We numerically observe that the error accuracy is less than 10^{-5} when $i = 10$.

V. Numerical results

In this section Monte Carlo simulations are carried out in order to evaluate the system performance of our proposed algorithm. To solve **Algorithm 1**, we use YALMIP toolbox [24] to model the optimization problem with SDPT3 [25] or MOSEK [26] as internal solver. For simulation, we consider the proportional fairness utility function, i.e., $f(\bar{r}_k) = \log(10^{-4} + \bar{r}_k)$ [29]. We denote our proposed user association algorithms for HetNet (resp. Homogeneous network) as **HetNet-Hybrid** (resp. **HomNet** [15]). Here, **HomNet** [15] refers to when the MBS serves both MUEs and SUEs without SCs. We compare our proposed algorithm with **HomNet** [15] and with the previous work [1] (**HetNet-Closed Access** [1]). **HetNet-Closed Access** [1] case considers only joint in-band scheduling and interference mitigation algorithm with fixed user association scheme (SCs are configured in closed subscriber group). The network performance are evaluated under the impact of the number of SCs per km^2, the number of MBS antennas N, and the MBS transmit power levels $P^{(b_0)}$ at low and high frequency bands. We provide the convergence behaviour of the proposed method and validation of the approximation method.

A. Simulation Environments

Consider a HetNet scenario, where a MBS is located at the center of a square area, MUEs are randomly deployed within the coverage of the MBS (the minimum MBS-MUE distance

TABLE I
PARAMETER SETTINGS

Path loss model [30]	Values in dB	Bandwidth (B)
LOS @ 28 GHz	$61.4 + 20\log(d)$	1 GHz
LOS @ 10 GHz	$55.25 + 18.5\log(d)$	100 MHz
LOS @ 2.4 GHz	$17 + 37.6\log(d)$	20 MHz
Parameter		Values
Maximum transmit power of MBS $P^{(b_0)}$		41 dBm
Channel quality τ		0.1
SC antenna gain		5 dBi
Lyapunov parameter ν		2×10^6

is 35 m). The SCs are uniformly distributed and one SUE per each SC is considered. The number of antennas at SCs N_s is greater than two, while we assume each SC can serve up to $N_s^{\text{au}} = 2$ UEs (including its own SUE). The path loss is modeled as a distance-based path loss with line-of-sight (LOS) model for urban environments [30]. We first assume that the probability of obtaining LOS is very high to make the performance evaluation, while the effect of other channel models is studied later. The FD interference threshold ϵ_o is set to 5×10^{-3} and the RZF parameter is $\alpha = 10^{-2}$. The data arrivals follow the Poisson distribution with the mean rate of 1 Gbps, 100 Mbps, and 20 Mbps for 28 GHz, 10 GHz, and 2.4 GHz, respectively. The parameter settings are summarized in Table I.

B. Ultra-Dense Small Cells Environment

To show the impact of network density, the average UE throughput (avgUT) and the cell-edge UE throughput (cell-edge UT) as a function of the number of SCs are shown in Fig. 3 and Fig. 4, respectively. The maximum transmit power of MBS and SCs is set to 41 dBm and 32 dBm, respectively[8]. In Fig. 3 and Fig. 4, the simulation is carried out in the asymptotic regime where the number of BS antennas and the network size (MUEs and SCs) grow large with a fixed ratio [7]. In particular, the number of SCs and the number of SUEs are both increased from 36 to 1000 per km^2, while the number of MUEs is scaled up with the number of SCs, such that $M = 1.5 \times S$. Moreover, the number of transmit antennas at MBS and SCs is set to $N = 2 \times K$ and $N_s = 6$, respectively. We recall that when adding SCs we also add one SUE per one SC that increases the network load. Here, the total number of users is increased while the maximum transmit power is fixed, and thus, the per-user transmit power is reduced with $1/K$, which reduces the per-UE throughput. Even though the number of MBS antennas is increased with K, as shown in Fig. 5 and Fig. 6, the performance of massive MIMO reaches the limit as the number of antennas goes to infinity. It can be seen that with increasing network load, our proposed algorithm **HetNet-Hybrid** outperforms baselines (with respect to the avgUT and the cell-edge UT) and the performance gap of the cell-edge UT is largest (5.6×) when the number of SC per km^2 is 350, and is small when the number of SC per km^2 is too small or too large. The reason is that when the number of SCs per km^2 is too small, the probability for an MUE to find a open access nearby-SC to connect is low. With increasing the number of SCs per km^2 MUEs are more likely to connect with open access nearby-SCs to increase the cell-edge UT. However, when the number of SCs per km^2 is too large, the cell-edge UT performance of **HetNet-Hybrid** is close to that of **HetNet-Closed Access** [1] due to the increased FD interference. Moreover, Fig. 3 and Fig. 4 show that the combination of Massive MIMO and FD-enabled SCs improves the network performance; for instance, **HetNet-Hybrid** and **HetNet-Closed Access** [1] outperform **HomNet** [15] in terms of both the avgUT and the cell-edge UT. Our results provide good insight for network deployment: for a given target UE throughput, what is the optimal number of UEs to schedule and what is the optimal/maximum number of SCs to be deployed?

[8]We reduce the maximum transmit power of MBS as compared to our previous work [1], which used that of 43 dBm. Hence, the average UT in this scenario is lower than [1].

C. Wireless Backhaul Impact versus Number of MBS Antennas

For a given number of UEs and SCs, we show the backhaul impact by varying the number of MBS antennas (MIMO gain). We also increase the number of SCs antennas, N_s, from 4 to 48. Here, we set the network area to 0.5 by 0.5 km^2, and consider 4 SCs and 8 MUEs. From the antenna theory [31], the beamforming gain is logarithmically proportional to the number of antennas, and thus, as the number of antennas goes to infinity, the beamforming gain diminishes. The avgUT and the cell-edge UT as a function of the number of MBS antennas are shown in Fig. 5 and Fig. 6, respectively. For a not-so-large number of MBS antennas, our proposed algorithm **HetNet-Hybrid** yields higher avgUT and cell-edge UT as compared to both baselines. For large number of antennas the MUEs choice of associating with the near-by SCs or the MBS yields similar payoffs, the gain of Massive MIMO by smart beamforming saturates. Hence, our proposed algorithm **HetNet-Hybrid** and **HetNet-Closed Access** [1] are tending to be the same as the number of antennas grows large. In Fig. 6, the performance of cell-edge UE throughput (cell-edge UT) of all schemes tends to be the same, when the number of antennas increases. Under the worst channel propagation of the cell-edge users, the performance of cell-edge users could not improve further, since all the network resources (transmit power and antennas) need to be shared among all UEs in order to maximize the total network utility.

D. Wireless Backhaul Impact versus Transmit Power Levels at different Frequency Bands

We also report the avgUT and the total network utility (TNU) along with the average queue length ("dashed line") as a function of the MBS maximum transmit power at different frequency bands (28 GHz, 10 GHz, and 2.4 GHz) in Fig. 7 and Fig. 8, respectively. In particular we consider the number of SCs is $S = 45$ per km^2, and the number of MUEs M is twice the number of SCs S. The number of MBS antennas is set to $N = K$, while the number of antennas at SCs $N_s + 1$ is set to 5. Due to insufficient number of antennas at the MBS to simultaneously serve all MUEs and SCs and to alleviate

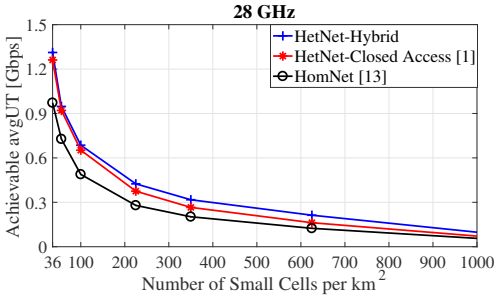

Fig. 3. Achievable avgUT versus number of small cells per km^2, S, when scaling $K = 2.5 \times S$, $N = 2 \times K$.

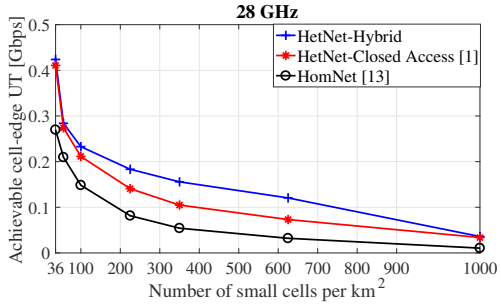

Fig. 4. Achievable cell-edge UT versus number of small cells per km^2, S, when scaling $K = 2.5 \times S$, $N = 2 \times K$.

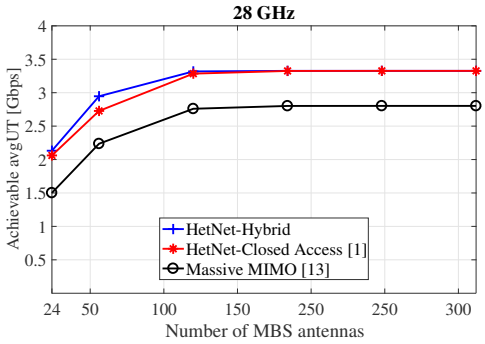

Fig. 5. Achievable avgUT versus N, when $K = 12$.

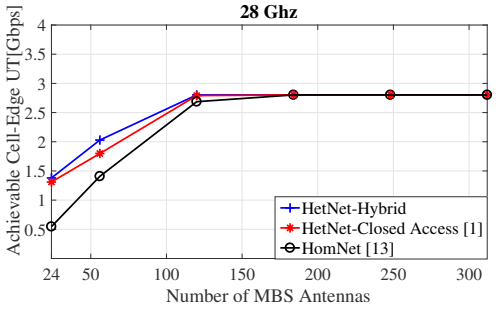

Fig. 6. Achievable cell-edge UT versus N, when $K = 12$.

the interference, offloading from the MBS to SCs helps to associate more UEs to the BSs. In this case the TNU is low, since the number of MBS antennas is reduced by half as compared to the impact of MBS antennas cases. As decreasing the maximum transmit power at the MBSs, **HetNet-Hybrid** outperforms **HetNet-Closed Access** [1], there is an inflexion point where the performance of **HetNet-Hybrid** is close to that of **HetNet-Closed Access** [1] when the transmit power level is 25 dBm, 31 dBm, and 37 dBm at 28 GHz, 10 GHz, and 2.4 GHz, respectively. It can be observed that at higher frequency bands FD-enabled SCs work better at open access mode than closed access mode under the same transmit power budget. When the maximum MBS transmit power is too small, the performance of **HetNet-Hybrid** and **HetNet-Closed Access** [1] is very closed to that of **HomNet** [15].

Moreover, in Fig. 9 we report the avgUT versus the ratio of number of MUEs to number of SCs, M/S, under different sets of SCs. Here, the number of SCs per km^2 is set to 45, 100, and 400 representing the network density from sparse to dense, while the ratio M/S is varying from 0.5 to 5. It can be observed that under the same total number of UEs, i.e., $K = 600$, deploying denser number of SCs with $S = 400$ and $M = 200$ obtains avgUT of 0.566 Gbps, which is higher than 0.3169 Gbps for a system with less number of SCs $S = 100$ and $M = 500$.

We have used the LOS channel model to make the performance evaluation such that the probability of obtaining LOS is very high. We now report the impact of channel models on massive MIMO system operating at 28 GHz mmWave frequency band. Beside the LOS and non-LOS (NLOS) channel states, there exists another channel state called blockage state, which is modeled as a distance-dependence probability state where the channel is either LOS or NLOS by using the stochastic model [32]. In Fig. 10, the performance gap between LOS and blockage channel models is shown versus the maximum transmit power.

E. Convergence

In Fig. 11 we show the convergence behaviour of our approximated algorithm based on the SCA method when deploying our **HetNet-Hybrid** algorithm. While the convergence analysis is provided in Appendix A. Unlike other works, we plot the cumulative distribution of the number of iterations at which the **Algorithm 1** converges for all t. We observe that the probability that the number of iterations takes on a value less than or equal to 4 is 90%, which implies that our proposed algorithm only needs few iterations to converge.

We then validate the accuracy of the closed-form expression for the data rate by comparing the Ergodic sum rate R, which is obtained by using the SINR from (7) and (8) from simulations of i.i.d. Rayleigh block-fading channels, to the approximated sum rate $\tilde{R}$, which obtained by using SINR from (13) and (14). The sum rate is defined as the total sum of all user data rates. We define the absolute error as $\frac{\tilde{R}-R}{\tilde{R}}$, then we plot the absolute error versus the number of MBS antennas, while the number of users is fixed to $K = 12$. As can be seen in Fig. 12, the absolute

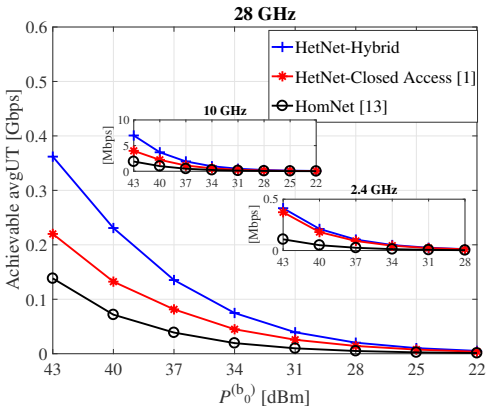

Fig. 7. Achievable avgUT versus $P^{(b_0)}$ at 28, 10, and 2.4 GHz, when $S = 45$ per km^2, $K = 3 \times S$, $N = K$.

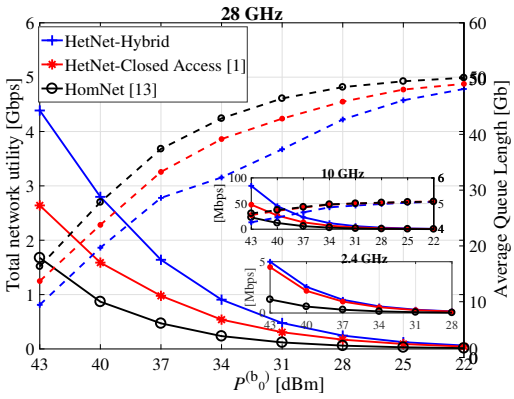

Fig. 8. TNU ("solid line") and network queue length ("dashed line") versus $P^{(b_0)}$ at 28, 10, and 2.4 GHz, when $S = 45$ per km^2, $K = 3 \times S$, $N = K$.

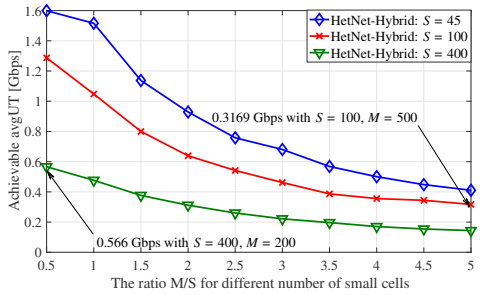

Fig. 9. Achievable avgUT versus the ratio M/S for different number of small cells.

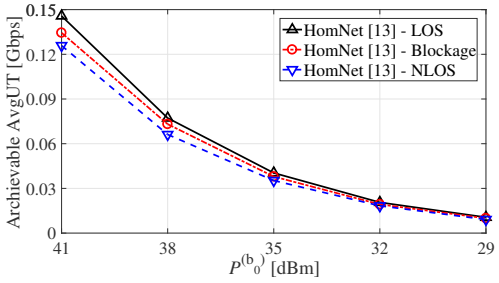

Fig. 10. Effects of LOS, NLOS, and blockage channel models, when $K = 3 \times S = 12$ and $N = K$.

error decreases as increasing number of MBS antennas. It means that closed-form expressions is more accurate when number of MBS antennas is higher than number of users, i.e., $N \gg K$. The impact of the Lyapunov parameter ν on the achievable average network utility and queue backlog has been showed in our previous work [1]. It has been observed that the network utility is increasing with $O(1/\nu)$, while the network backlog linearly increases with $O(\nu)$. Hence, choosing the value of ν will result in an $[O(1/\nu), O(\nu)]$ utility-queue backlog tradeoff, which leads to an utility-delay tradeoff [11].

F. Effect of Pilot Training and Channel Aging

In this subsection we study the effect of pilot training and channel aging in 5G Massive MIMO system as we consider a large number of antennas and users. We assume that each coherence interval consists of three phases: uplink (UL) training, DL payload data transmission, and UL payload data transmission. During the UL training phase, the users send pilot sequences to the BSs and each BS estimates the channels. The estimate channels are used to precode the transmit signals in the DL. Let T_{ci} and τ_{td} denote the length of the coherence interval and the UL training duration, where the subscripts ci and td stand for "coherence interval" and "training duration", respectively. Basically, we have the length of pilot sequences is less than the length of the coherence interval, i.e., $\tau_{\text{td}} < T_{\text{ci}}$. If the length of pilot sequences τ_{td} is greater or equal to the number of user K, $\tau_{\text{td}} \geq K$, to achieve the estimate channel the pilot assignment is chosen pairwisely orthogonal. However, if $\tau_{\text{td}} < K$, then channel estimate is degraded due to non-orthogonal pilot signals that leads to the pilot contamination effect. In this work we consider the imperfect CSI, which is due to channel estimation errors during the UL training and the coherence interval T_{ci}, while the channel reciprocity is perfect for UL and DL. In addition, the channel aging is also a very important issue needed to be addressed in Massive MIMO systems, and the Massive MIMO systems are most suitable for static users with not-too-fast movement. For simplicity, we consider the problem of channel aging by the impact of channel estimate error factor τ_k as shown in (1). We next show the relation between the channel estimate error and the length of pilot sequences. Assume that each user will transmit the orthogonal pilot signal to the MBS during the UL training in which $\tau_{\text{td}} \geq K$ and the MBS receives the pilot signals simultaneously. Follow the analysis in [15], the channel estimate error is $\tau_k^2 = \frac{1}{1+\tau_{\text{td}}\rho^{\text{ul}}}$, where ρ^{ul} is defined as the UL signal-to-noise ratio of user k, $\rho^{\text{ul}} = p_k^{\text{ul}}/\sigma_k^2$, here p_k^{ul} is the UL transmit power of user k.

Since our work considers only the performance of the DL, we then simplistically assume that the transmission time for DL and UL during the coherence time is identical[9]. Therefore, the closed-form expression of UT taking into account the

[9]Due to different traffic model for DL and UL, the ratio configuration for DL and UL transmission time will be varied.

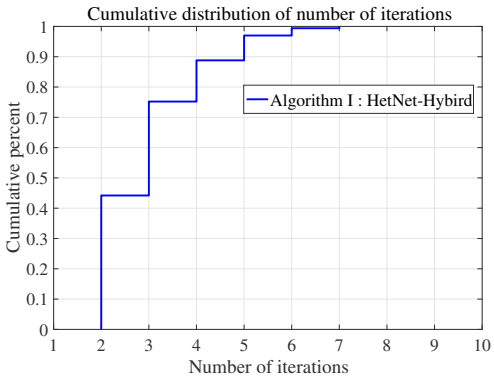

Fig. 11. The CDF of convergence of **Algorithm 1**.

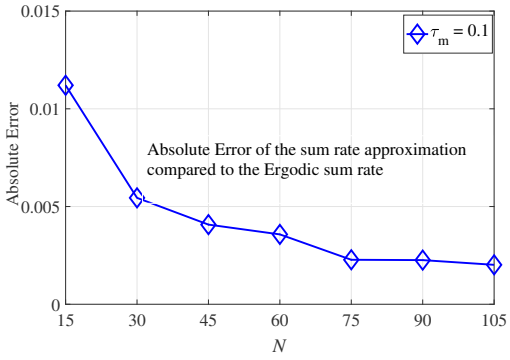

Fig. 12. Validation of the approximation of the closed-form expression, when $K = 12$ and $K = 3 \times S$.

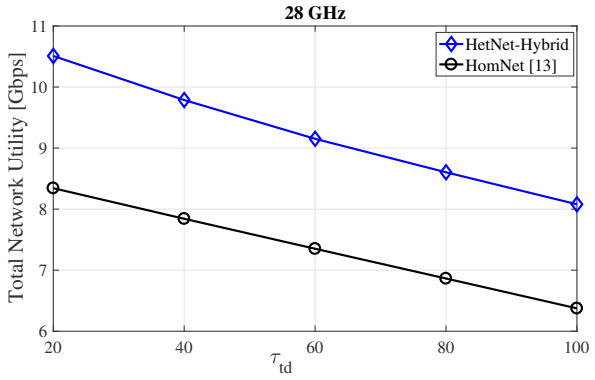

Fig. 13. Impact of pilot sequence length, when $K = 12$, $N = 2 \times K$, and $K = 3 \times S$.

impact of pilot training, coherence interval, and channel aging is

$$r_k^{\text{ci}}(\mathbf{\Lambda}|\mathbf{\Theta}) \xrightarrow{a.s.} \text{B}\frac{(1-\tau_{\text{td}}/T_{\text{ci}})}{2}\log\left(1+\frac{l_k^{(b_0)}p_k^{(b_0)}(1-\frac{1}{1+\tau_{\text{td}}\rho^{\text{ul}}})}{1+\sum_{s=1}^{S}\beta^{(b_s)}\zeta_k^{(b_s)}}\right),$$

where B is the channel bandwidth. To set up the simulation parameters for this subsection all the transmit powers are normalized by dividing by the thermal noise power σ_k. At 28 GHz we set the coherence interval to $T_{\text{ci}} = 350$ with the coherence bandwidth of 10 MHz and the coherence time of 35 µs.

We show the impact of pilot training on the total network utility of our proposed HetNet-Hybrid as compared to HomNet at 28 GHz. To do that, we vary the length of pilot sequences from 20 to 100, while the number of pilot training sequence is greater than number of users. As can be seen in Fig. 13, with increasing the pilot training duration, the total network utility for the DL is gradually degraded.

VI. Conclusion

We have studied the NUM by considering the problem of joint load balancing and interference mitigation taking into account the dynamic wireless backhaul, traffic demand, and imperfect CSI in 5G HetNets. The problem of load balancing has taken into account the wireless backhaul capacity in order to reduce the load of the MBS while avoiding the bottleneck problem at the FD-enabled SCs. By utilizing a very large number of antennas at MBS we design a hierarchical precoder in order to mitigate both cross-tier and co-tier interference in HetNets. We aim to maximize a network utility function of the total time-average data rates subject to the backhaul dynamic and network stability in the presence of imperfect CSI. We have exploited from RMT to obtain a closed-form expression of the original problem in a large system regime. By applying the stochastic optimization, the problem is then decoupled decoupled into dynamic scheduling of MUEs, backhaul provisioning of in-band FD-enabled SCs, and offloading UEs to in-band FD-enabled SCs as a function of interference, number of antennas, and backhaul loads. We have provided the performance analysis and validation of our proposed algorithm, which states that there exists an $[O(1/\nu), O(\nu)]$ utility-queue backlog tradeoff. Via numerical results, we have show that our proposed algorithm outperforms the baselines with respect to the number of SCs per km^2, the number of MBS antennas, and the MBS transmit power levels at different frequency bands. Interestingly, we find that even at lower frequency band the performance of open access small cell is close to that of closed access at some operating points, the open access full-duplex small cell still yields higher gain as compared to the closed access at higher frequency bands. With increasing the small cell density or the wireless backhaul quality, the open access full-duplex small cells outperform and achieve 5.6× gain in terms of cell-edge performance as compared to the closed access ones in ultra-dense networks with 350 small cell base stations per km^2.

Appendix A
Convergence Analysis for **Algorithm 1**

Next, we establish a convergence result for **Algorithm 1** based on the SCA method, since the original problem (22) has a non-convex objective function (22a) subject to non-convex constraint (22f). By using the SCA method, we replace the original nonconvex problem (22) by a strongly convex problem (27). We will briefly describe the convergence here

for the sake of completeness since it was studied in [21], [22]. We assume that the **Algorithm 1** obtains the solution of problem (27) at iteration $i+1$ th. The updating rule in **Algorithm 1** ensures that the optimal values $\mathbf{\Lambda}^{o(i)}$, $\lambda_{ks}^{(i)}$, and $\hat{\iota}_s^{(i)}$ at iteration i satisfy all constraints in (27) and are feasible to the optimization problem at iteration $i+1$. Hence, the objective obtained in the $i+1$st iteration is less than or equal to that in the in the ith iteration, since we minimize the convex function. In other words, **Algorithm 1** yields a non-increasing sequence. Due to antenna and interference constraints, the objective is bounded, and thus **Algorithm 1** converges to some local optimal solution of (27). Moreover, **Algorithm 1** produces a sequence of points that are feasible for the original problem (22) and this solution is satisfied the KKT condition of the original problem (22) as discussed in [21], [22].

Appendix B
Performance Analysis

Theorem 3 is provided to show the performance analysis of network utility maximization based on Lyapunov framework and prove that the queues are stable.

Theorem 3: [Optimality] Assume that all queues are initially empty. For arbitrary arrival rates, the operation mode and load balancing is chosen to satisfy (21) and the rate regime. For a given constant $\chi \geq 0$, the network utility maximization with any $\nu > 0$ provides the following utility performance with $\chi - approximation$

$$f_0 \geq f_0^* - \frac{\Psi + \chi}{\nu},$$

where $f_0^\star$ is the optimal network utility over the rate regime. *Proof:* To prove the **Theorem 3**, we first prove the queues are bounded. Let π_k denote the largest right derivative of $f(\bar{r}_k)$, the Lyapunov framework can guarantee the following strong stability of the virtual queues and the network queues.

$$Q_k(t) \leq \nu\omega_k(t)\pi_k + 2a_k^{\max}, \tag{34}$$

$$Y_k(t) \leq \nu\omega_k(t)\pi_k + a_k^{\max}, \tag{35}$$

$$D_s(t) \leq \nu\omega_s(t)\pi_s + a_{s+M}^{\max}. \tag{36}$$

Here we first prove the bound of the virtual queues, and then the bound of the network queues are proved similarly. Suppose that all queues are initially empty, this clearly holds for $t = 0$. Suppose these inequalities hold for some $t > 0$, we need to show that it also holds for $t+1$.

From (17) and (19), if $Y_k(t) \leq \nu\omega_k(t)\pi_k$ and $D_s(t) \leq \nu\omega_s(t)\pi_s$ then $Y_k(t+1) \leq \nu\omega_k(t)\pi_k + a_k^{\max}$ and $D_s(t+1) \leq \nu\omega_s(t)\pi_s + a_{s+M}^{\max}$ and the bound holds for $t+1$ due to the arrival rate constraint $r_k(t) \leq a_k^{\max}$ and $r_s(t) \leq a_s^{\max}$. Else, if $Y_k(t) \geq \nu\omega_k(t)\pi_k$ and $D_s(t) \geq \nu\omega_s(t)\pi_s$; since the value of auxiliary variables is determined by maximized $\sum_{k=1}^K Y_k(t)\varphi_k(t) + \sum_{s=1}^S D_s(t)\varphi_{s+M}(t) - \nu f_0(\boldsymbol{\varphi}(t))$, $\boldsymbol{\varphi}(t)$ is then forced to be zero. From (19) and (17), $Y_k(t+1)$ and $D_s(t+1)$ are bounded by $Y_k(t)$ and $D_s(t)$, respectively. Since the virtual queues are bounded for t, we have the following inequalities

$$Y_k(t+1) \leq Y_k(t) \leq \nu\omega_k(t)\pi_k + a_k^{\max}, \tag{37}$$

$$D_s(t+1) \leq D_s(t) \leq \nu\omega_s(t)\pi_s + a_{s+M}^{\max}. \tag{38}$$

Hence, the bounds of the virtual queues hold for all t. Similarly, we show that the network queue (34) holds for all t. It clearly holds for $t = 0$. We assume that (34) holds for $t > 0$, we now prove it holds for $t+1$. Note that from (16) and (19) we have $Q_k(t+1) \leq H_k(t+1) + a_k(t)$. Moreover, we just proved that $H_k(t+1) \leq \nu\omega_k(t)\pi_k + a_k^{\max}$ then we have $Q_k(t+1) \leq \nu\omega_k(t)\pi_k + 2a_k^{\max}$ and the network bound holds for $t+1$.

We have established the network bounds, we are going to show the utility bound. Since our solution of (18) is to minimize the Lyapunov drift and the objective function every time slot t, we have the following inequality

$$\begin{aligned}\mathbf{\Delta}(\Sigma(t)) - \nu\mathbb{E}[f_0(\boldsymbol{\varphi}(t))] \leq \\ \Psi - \nu\mathbb{E}[f_0(\boldsymbol{\varphi}^*(t))] + \textstyle\sum_{k=1}^K Q_k(t)\mathbb{E}\Big[a_k(t) - r_k^*(t)|\Sigma(t)\Big] \\ + \textstyle\sum_{k=1}^K Y_k(t)\mathbb{E}\Big[\varphi_k^*(t) - r_k^*(t)|\Sigma(t)\Big] \\ + \textstyle\sum_{s=1}^S D_s(t)\mathbb{E}\Big[\varphi_{s+M}^*(t) - \beta^{(b_s)*}(t)r_s^{c_s*}(t)|\Sigma(t)\Big],\end{aligned}$$

where $\boldsymbol{\varphi}^*(t), \beta^{(b_s)*}(t)$, and $r_k^*(t)$ are the optimal values of the problem (21). Since the queues are bounded, for given $\chi \geq 0$, obtaining

$$\mathbf{\Delta}(\Sigma(t)) - \nu\mathbb{E}[f_0(\boldsymbol{\varphi}(t))] \leq \Psi - \nu\mathbb{E}[f_0(\boldsymbol{\varphi}^*(t))] + \chi.$$

By taking expectations of both sides of the above inequality and choosing $\mathbf{r}^*(t) = \boldsymbol{\varphi}^*(t)$, it yields for all $t \geq 0$,

$$\begin{aligned}\mathbb{E}\Big[L(\Sigma(t+1)) - L(\Sigma(t))|\Sigma(t)\Big] - \nu\mathbb{E}[f_0(\boldsymbol{\varphi}(t))] \leq \\ \Psi + \chi - \nu\mathbb{E}[f_0(\mathbf{r}^*(t))].\end{aligned}$$

By taking the sum over $\tau = 0, \ldots, t-1$ and dividing by t, (using the fact that $f_0(\mathbf{r}^*(t)) = f_0^*$), yielding

$$\begin{aligned}\frac{\mathbb{E}\Big[L(\Sigma(t+1)) - L(\Sigma(0))|\Sigma(t)\Big]}{t} - \frac{\nu}{t}\sum_{\tau=0}^{t-1}\mathbb{E}[f_0(\boldsymbol{\varphi}(t))] \leq \\ \Psi + \chi - \nu f_0^*.\end{aligned} \tag{39}$$

By using the fact that $L(\Sigma(t+1)) \geq 0$ and $L(\Sigma(0)) = 0$, and applying Jensen's inequality in the concave function and rearranging term yields

$$f_0(\boldsymbol{\varphi}(t)) \geq \ f_0^* - \frac{\Psi + \chi}{\nu}.$$

Since the network utility function is a non-decreasing concave function, the auxiliary variable is chosen to satisfy $r_k(t) \geq \varphi_k(t)$. Hence $f_0(\mathbf{r}(t)) \geq f_0(\boldsymbol{\varphi}(t)) \geq \ f_0^* - \frac{\Psi+\chi}{\nu}$, which means that the solution is closed to the optimal as increasing ν. Which completes the proof of the **Theorem 3**. Hence, there exists an $[O(\nu), O(1/\nu)]$ utility-queue length tradeoff, which leads to an utility-delay balancing.

We now prove that all queues are stable by using the **Definition 3**, the bound (39) can be rewritten as

$$\mathbf{\Delta}(\Sigma(t)) \leq \ C,$$

where C is any constant that satisfies for all t and $\Sigma(t)$: $C \geq \Psi + \chi - \nu(f_0^* - \mathbb{E}[f_0(\boldsymbol{\varphi}(t))])$. By using the definition of the Lyapunov drift and taking an expectation, obtaining

$$\mathbb{E}\Big[L(\Sigma(t))\Big] \leq \ Ct.$$

As the definition of the Lyapunov function $L(\Sigma(t))$ we have

$$\mathbb{E}[Q_k(t)]^2, \mathbb{E}[H_k(t)]^2, \mathbb{E}[D_s(t)]^2 \leq 2Ct.$$

Dividing both sides by t^2, and taking the square roots shows for all $t > 0$:

$$\frac{\mathbb{E}[Q_k(t)]}{t}, \frac{\mathbb{E}[H_k(t)]}{t}, \frac{\mathbb{E}[D_k(t)]}{t} \leq \sqrt{\frac{2C}{t}}.$$

As $t \to \infty$, taking the limit, we prove the queues are stable.

ACKNOWLEDGMENT

The authors would like to acknowledge colleagues: Kien Giang Nguyen, Mohammed ElBamby, and Petri Luoto for helpful discussions on the paper.

REFERENCES

[1] T. K. Vu, M. Bennis, S. Samarakoon, M. Debbah, and M. Latva-aho, "Joint In-Band Backhauling and Interference Mitigation in 5G Heterogeneous Networks," in *Proceedings of 22th European Wireless Conference*, Oulu, Finland, May 2016, pp. 1–6.

[2] Nokia Siemens Networks, "2020: Beyond 4G Radio Evolution for the Gigabit Experience," White Paper, Nokia Siemens Networks, 2011.

[3] T. L. Marzetta, "Noncooperative cellular wireless with unlimited numbers of base station antennas," *IEEE Transactions on Wireless Communications*, vol. 9, no. 11, pp. 3590–3600, 2010.

[4] T. K. Vu, S. Kwon, and S. Oh, "Cooperative Interference Mitigation Algorithm in Heterogeneous Networks," *IEICE Transactions on Communications*, vol. 98, no. 11, pp. 2238–2247, 2015.

[5] L. Boyu et al., "Small cell in-band wireless backhaul in massive Multiple-Input Multiple-Output systems," in *IEEE International Conference on Communications*, London, UK, June 2015, pp. 1838–1844.

[6] S. Hur et al., "Millimeter wave beamforming for wireless backhaul and access in small cell networks," *IEEE Transactions on Communications*, vol. 61, no. 10, pp. 4391–4403, 2013.

[7] L. Sanguinetti, A. Moustakas, and M. Debbah, "Interference management in 5G reverse TDD HetNets: A large system analysis," *IEEE Journal on Selected Areas in Communications*, vol. 33, pp. 1187–1200, 2015.

[8] Q. Ye et al., "User association for load balancing in heterogeneous cellular networks," *IEEE Transactions on Wireless Communications*, vol. 12, no. 6, pp. 2706–2716, 2013.

[9] D. Bethanabhotla et al., "Optimal User-Cell Association for Massive MIMO Wireless Networks," *IEEE Transactions on Wireless Communications*, vol. 15, no. 3, pp. 1835–1850, March 2016.

[10] D. Liu, L. Wang, Y. Chen, M. Elkashlan, K. K. Wong, R. Schobe, and L. Hanzo, "User Association in 5G Networks: A Survey and an Outlook," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1018–1044, 2016.

[11] T. K. Vu et al., "Ultra-Reliable and Low Latency Communication in mmWave-Enabled Massive MIMO Networks," *IEEE Communications Letters*, vol. 21, no. 9, pp. 1–4, 2017.

[12] N. Omidvar et al., "Optimal Hierarchical Radio Resource Management for HetNets with Flexible Backhaul," *submitted to IEEE Transactions on Signaling Processing*, 2016.

[13] J. Andrews, S. Singh, Q. Ye, X. Lin, and H. Dhillon, "An overview of load balancing in HetNets: Old myths and open problems," *IEEE Wireless Communications*, vol. 21, no. 2, pp. 18–25, 2014.

[14] F. Rusek et al., "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Processing Magazine*, vol. 30, no. 1, pp. 40–60, 2013.

[15] S. Wagner et al., "Large system analysis of linear precoding in correlated MISO broadcast channels under limited feedback," *IEEE Transactions on Information Theory*, vol. 58, no. 7, pp. 4509–4537, 2012.

[16] A. Liu and V. Lau, "Hierarchical Interference Mitigation for Massive MIMO Cellular Networks," *IEEE Transactions on Signal Processing*, vol. 62, no. 18, pp. 4786–4797, Sept 2014.

[17] Z. Jun et al., "Large system analysis of cognitive radio network via partially-projected regularized zero-forcing precoding," *IEEE Transactions on Wireless Communications*, vol. 14, no. 9, pp. 4934–4947, 2015.

[18] H. Boche, S. Naik, and M. Schubert, "Pareto boundary of utility sets for multiuser wireless systems," *IEEE/ACM Transactions on Networking*, vol. 19, no. 2, pp. 589–601, 2011.

[19] Z. Chen et al., "Pareto region characterization for rate control in MIMO interference systems and Nash bargaining," *IEEE Transactions on Automatic Control*, vol. 57, no. 12, pp. 3203–3208, 2012.

[20] M. J. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synthesis Lectures on Communication Networks*, vol. 3, no. 1, pp. 1–211, 2010.

[21] A. Beck, A. Ben-Tal, and L. Tetruashvili, "A sequential parametric convex approximation method with applications to nonconvex truss topology design problems," *Journal of Global Optimization*, vol. 47, no. 1, pp. 29–51, 2010.

[22] L. N. Tran et al., "Fast converging algorithm for weighted sum rate maximization in multicell MISO downlink," *IEEE Signal Processing Letters*, vol. 19, no. 12, pp. 872–875, 2012.

[23] H. Li, L. Song, and M. Debbah, "Energy efficiency of large-scale multiple antenna systems with transmit antenna selection," *IEEE Transactions on Communications*, vol. 62, no. 2, pp. 638–647, 2014.

[24] J. Löfberg, "YALMIP: A toolbox for modeling and optimization in MATLAB," in *IEEE International Symposium on Computer Aided Control Systems Design*, New Orleans, LA, USA, 2004, pp. 284–289.

[25] K.-C. Toh, M. J. Todd, and R. H. Tütüncü, "SDPT3 - a MATLAB software package for semidefinite programming, version 1.3," *Optimization methods and software*, vol. 11, no. 1-4, pp. 545–581, 1999.

[26] A. MOSEK, "The MOSEK optimization toolbox for MATLAB manual, Version 7.1 (Revision 28)," *http://mosek. com,(accessed on March 20, 2015)*, 2015.

[27] A. Ben-Tal and A. Nemirovski, "On polyhedral approximations of the second-order cone," *Mathematics of Operations Research*, vol. 26, no. 2, pp. 193–205, 2001.

[28] K.-G. Nguyen, L.-N. Tran, O. Tervo, Q.-D. Vu, and M. Juntti, "Achieving energy efficiency fairness in multicell MISO downlink," *IEEE Communications Letters*, vol. 19, no. 8, pp. 1426–1429, 2015.

[29] J. Mo and J. Walrand, "Fair end-to-end window-based congestion control," *IEEE/ACM Transactions on Networking (ToN)*, vol. 8, no. 5, pp. 556–567, 2000.

[30] A. Mustafa Riza et al., "Millimeter wave channel modeling and cellular capacity evaluation," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 6, pp. 1164–1179, 2014.

[31] C. A. Balanis, *Antenna theory: analysis and design*. John Wiley & Sons, 2005.

[32] T. Bai et al., "Millimeter wave cellular channel models for system evaluation," in *IEEE International Conference on Computing, Networking and Communications*, Honolulu, HI, USA, April 2014, pp. 178–182.

PLACE PHOTO HERE

Trung Kien Vu received the B.Eng. degree from the School of Electronics and Telecommunications, Hanoi University of Science and Technology, Vietnam, in 2012, and the M.Sc. degree in electrical engineering from the School of Electrical Engineering, University of Ulsan, South Korea, in 2014. He is currently pursuing the D.Sc. degree with the Centre for Wireless Communications, University of Oulu, Finland. His research interests focus on heterogeneous cellular networks, massive MIMO, mm-wave communications, network planning and optimization, and ultra-reliable and low latency communications. He received the 2016 European Wireless Best Paper Award and was a recipient of the 2014 Brain Korean 21 Plus (BK21+) Scholarship.

Mehdi Bennis received his M.Sc. degree in Electrical Engineering jointly from the EPFL, Switzerland and the Eurecom Institute, France in 2002. From 2002 to 2004, he worked as a research engineer at IMRA-EUROPE investigating adaptive equalization algorithms for mobile digital TV. In 2004, he joined the Centre for Wireless Communications (CWC) at the University of Oulu, Finland as a research scientist. In 2008, he was a visiting researcher at the Alcatel-Lucent chair on flexible radio, SUPELEC. He obtained his Ph.D. in December 2009 on spectrum sharing for future mobile cellular systems. Currently Dr. Bennis is an Adjunct Professor at the University of Oulu and Academy of Finland research fellow. His main research interests are in radio resource management, heterogeneous networks, game theory and machine learning in 5G networks and beyond. He has co-authored one book and published more than 100 research papers in international conferences, journals and book chapters. He was the recipient of the prestigious 2015 Fred W. Ellersick Prize from the IEEE Communications Society, the 2016 Best Tutorial Prize from the IEEE Communications Society and the 2017 EURASIP Best paper Award for the Journal of Wireless Communications and Networks. Dr. Bennis serves as an editor for the IEEE Transactions on Wireless Communication.

Sumudu Samarakoon received his B. Sc. (Hons.) degree in Electronic and Telecommunication Engineering from the University of Moratuwa, Sri Lanka and the M. Eng. degree from the Asian Institute of Technology, Thailand in 2009 and 2011, respectively. He is currently persuading Dr. Tech degree in Communications Engineering in University of Oulu, Finland. Sumudu is also a member of the research staff of the Centre for Wireless Communications (CWC), Oulu, Finland. His main research interests are in heterogeneous networks, radio resource management, machine learning, and game theory.

Mérrouane Debbah entered the Ecole Normale Suprieure Paris-Saclay (France) in 1996 where he received his M.Sc. and Ph.D. degrees respectively. He worked for Motorola Labs (Saclay, France) from 1999-2002 and the Vienna Research Center for Telecommunications (Vienna, Austria) until 2003. From 2003 to 2007, he joined the Mobile Communications department of the Institut Eurecom (Sophia Antipolis, France) as an Assistant Professor. Since 2007, he is a Full Professor at CentraleSupelec (Gif-sur-Yvette, France). From 2007 to 2014, he was the director of the Alcatel-Lucent Chair on Flexible Radio. Since 2014, he is Vice-President of the Huawei France R&D center and director of the Mathematical and Algorithmic Sciences Lab. His research interests lie in fundamental mathematics, algorithms, statistics, information & communication sciences research. He is an Associate Editor in Chief of the journal Random Matrix: Theory and Applications and was an associate and senior area editor for IEEE Transactions on Signal Processing respectively in 2011-2013 and 2013-2014. Mérrouane Debbah is a recipient of the ERC grant MORE (Advanced Mathematical Tools for Complex Network Engineering). He is a IEEE Fellow, a WWRF Fellow and a member of the academic senate of Paris-Saclay. He has managed 8 EU projects and more than 24 national and international projects. He received 17 best paper awards, among which the 2007 IEEE GLOBECOM best paper award, the Wi-Opt 2009 best paper award, the 2010 Newcom++ best paper award, the WUN CogCom Best Paper 2012 and 2013 Award, the 2014 WCNC best paper award, the 2015 ICC best paper award, the 2015 IEEE Communications Society Leonard G. Abraham Prize, the 2015 IEEE Communications Society Fred W. Ellersick Prize, the 2016 IEEE Communications Society Best Tutorial paper award, the 2016 European Wireless Best Paper Award and the 2017 Eurasip Best Paper Award as well as the Valuetools 2007, Valuetools 2008, CrownCom2009, Valuetools 2012 and SAM 2014 best student paper awards. He is the recipient of the Mario Boella award in 2005, the IEEE Glavieux Prize Award in 2011 and the Qualcomm Innovation Prize Award in 2012.

Matti Latva-aho received the M.Sc., Lic.Tech. and Dr. Tech (Hons.) degrees in Electrical Engineering from the University of Oulu, Finland in 1992, 1996 and 1998, respectively. From 1992 to 1993, he was a Research Engineer at Nokia Mobile Phones, Oulu, Finland after which he joined Centre for Wireless Communications (CWC) at the University of Oulu. Prof. Latva-aho was Director of CWC during the years 1998-2006 and Head of Department for Communication Engineering until August 2014. Currently he is Professor of Digital Transmission Techniques at the University of Oulu. He serves as Academy of Finland Professor in 2017-2022. His research interests are related to mobile broadband communication systems and currently his group focuses on 5G and beyond systems research. Prof. Latva-aho has published 300+ conference or journal papers in the field of wireless communications. He received Nokia Foundation Award in 2015 for his achievements in mobile communications research.