Energy-Efficient Beam Scheduling for Orthogonal Random Beamforming in Cooperative Networks

Jaehwan Jeong†, Jeongho Kwak‡ and Song Chong†
†School of Electrical Engineering, KAIST
‡INRS-EMT & Western University
E-mail: jh.jeong@netsys.kaist.ac.kr, jhkwak.inrs@gmail.com, songchong@kaist.edu

***Abstract*—In this paper, we study a joint beam and user scheduling problem in a cooperative cellular network utilizing orthogonal random beamforming technique. This paper aims to minimize total base stations' average energy expenditure while ensuring finite service time for all traffic arrivals in a given set. We leverage Lyapunov optimization technique to transform original long-term problem into short-term modified max-weight problem without knowledge of future network states such as traffic arrivals. We introduce a parameter which manipulates energy-delay tradeoff in our system as well. Since provided short-term problem is combinatorial and nonlinear optimization problem, we are inspired by a greedy algorithm to design near-optimal joint beam and user scheduling policy, namely BEANS. We prove that proposed BEANS *(i)* ensures finite service time for all traffic arrival rates within close to 1/2 capacity region and all (energy-delay) tradeoff parameters thanks to submodular characteristics of the objective function, and *(ii)* attains finite upper bounds of average energy consumption and average queue backlog for all traffic arrival rates within close to 1/4 capacity region and all tradeoff parameters. Finally, via extensive simulations, we compare the capacity region and energy-queue backlog tradeoff of BEANS with optimal and existing algorithms, and show that BEANS attains 43% of energy saving for the same average queue backlog compared to the algorithms which do not take traffic dynamics and energy consumption into considerations.**

I. Introduction

Motivation. We witness an explosion of mobile data traffic over the past years and this trend is expected to continue in the coming years[1]. This unprecedented usage of mobile data leads to tremendous amount of energy expenditure of network infrastructure[2]. In particular, energy consumption of base stations (BSs) is known to account for more than 80% of total energy consumption in cellular networks [3], which implies that BS energy saving exerts a great influence on entire network greening.

To date, the BS energy saving has been extensively studied in domains of various time scales [4]–[7]. Kwak *et al.* [4] suggested network-wide BS power sharing policies in a time scale of transmit power control, say few msec, for given total BSs' energy budget. Oh *et al.* [5] and Abbasi *et al.* [6] addressed BS activation policy in the presence of spatio-temporal traffic dynamics with a time scale of few hours. Son *et al.* [7] suggested joint user association and BS activation policy by leveraging a time scale separation of microscopic flow-level dynamics (e.g., few minutes) and macroscopic traffic-level dynamics (e.g., few hours).

Theoretically, wireless capacity can be linearly improved as the number of transmit and receive antennas increases [8]. In addition, radiated energy can be saved by focusing energy into ever-smaller regions of space with a beam formed by several transmit and receive antennas. Meanwhile, a recent trend on hyper-dense heterogeneous and small cell deployment to cope with pronounced increment of data traffic brings a severe inter-cell interference problem [9]. CoMP (Coordinated MultiPoint) technology [10] helps the network system cancel the inter-cell interference suffered from transmission of neighboring BSs by a cooperation of data transmission among BSs. Hence, the both MIMO (Multiple Input Multiple Output) and CoMP technologies allow the network system to attain higher data rate as well as higher energy saving for the same data rate performance. However, existing studies in the MIMO CoMP cellular system have mainly focused on the design of physical layer such as hybrid beamforming [11] or antenna precoding [12] whereas the network management policies such as a joint beam and user scheduling in the same network system were difficult to be addressed[3] due to high computational complexity of intertwined design with the antenna precoding[4].

Summary and contribution. In this paper, we formulate a joint beam and user scheduling problem in multi-cell cooperative cellular networks aiming to minimize total energy expenditure of base stations while ensuring finite service time with respect to dynamic traffic arrivals. We take into account an orthogonal random beamforming scheme as a precoding vector design [14] so as to mainly focus on the joint beam and user scheduling policy per se. The random beamforming does not require channel state information (CSI) of all channels between all transmit-receive antenna pairs, and only demand SINR (signal to interference plus noise ratio) feedbacks of all users, hence it significantly reduces computational complexity. Then, each user is matched to one of randomly generated

[1]According to the forecast of Cisco [1], global mobile data traffic will increase nearly eightfold between 2015 and 2020.

[2]According to the SMART 2020 report [2], the energy consumption of network infrastructure is expected to triple in 2020 compared to 2002.

[3]Some of recent work (e.g., see [10]) have addressed network management such as BS sleeping control for a given user scheduling in the MIMO CoMP system.

[4]Shi *et al.* [13] proposed a joint design of RRH (Remote Radio Head) (or BS) selection and coordinated transmit beamforming in C-RAN architecture to minimize network energy consumption assuming perfect channel estimation which is difficult to realize in practice.

978-3-901882-90-6/17 ©2017 IFIP

beamforming vectors so as to maximize their own objectives[5].

A prediction of traffic variations (channel variations as well) is difficult to realize because the time scale of beam/user scheduling problem is much shorter, say few msec, even though existing BS activation work (e.g., see [5]) assumes that they are predictable due to much longer control time scales, say several hours[6]. Hence, in the light of the prediction difficulities, we design an online-fashioned joint beam and user scheduling algorithm, i.e., selecting several pairs of a randomly generated beamforming vector and a user to be scheduled every time sequence, without knowledge of future network dynamics such as wireless states and traffic arrivals.

Towards this end, we invoke the Lyapunov optimization method [16] for which the short-term modified max-weight problem derived from the Lyapunov drift must be optimized in the slot-by-slot basis. In this short-term problem, we introduce a parameter which manipulates enegy-delay tradeoff in our system, i.e., total BSs' energy expenditure can be further saved by trading extra queue backlog with larger (energy-delay) tradeoff parameter[7].

In addition, because the modified combinatorial scheduling problem is NP-hard and difficult to tackle [18], we propose an approximation algorithm, namely BEANS, which eventually ensures minimum average BS energy expenditure and finite service time for all traffic arrivals within close to 1/2 capacity region characterized by a set of average arrival rates of which our system can serve within finite time for all tradeoff parameters. Going a step further, we prove the BEANS attains finite upper bounds of average energy expenditure and average queue backlog for all traffic arrival rates within close to 1/4 capacity region and all tradeoff parameters.

We should be noted that one of previous BS activation studies in a single antenna system [6] applied Lyapunov optimization as well to design their policy, but they did not show the performance bounds of proposed algorithm in perspectives of energy and queue backlog. Indeed, technical contribution of this paper is first to deliver a demonstration of energy and queue bounds of BEANS algorithm even though the short-term problem of our system is made of *submodular max-weight style function with negative terms* which are known to be very challenging to develop algorithms which guarantee the performance bounds[8]. Finally, via extensive simulations, we compare energy-delay tradeoff and queue stability region of BEANS with optimal and existing algorithms. Main contributions of this paper can be summarized as follows.

- We formulate a joint beam and user scheduling problem in a multi-cell cooperative networks with respect to spatio-temporal traffic dynamics aiming to minimize total BSs' average energy consumption while ensuring finite service time for all traffic arrivals within capacity region.
- We propose a uncomplicated joint beam and user scheduling algorithm, namely BEANS which guarantees a constant-ratio lower bound of an optimal performance. BEANS is pragmatic as well because it does not demand challenging prediction of traffic arrival and wireless channel states.
- We prove that BEANS ensures finite service time for all arrival rates within close to 1/2 capacity region and attains upper bounds of total BSs' average energy expenditure and average queue backlog for all arrival rates within close to 1/4 capacity region and all tradeoff parameters by leveraging submodular characteristics of short-term problem.
- Via extensive simulations, we show that BEANS saves 43% of average BS energy for the same average queue backlog, and reduces 50% of average queue backlog for the same energy consumption compared to baseline algorithms which do not take energy consumption and queue backlog into considerations. In addition, BEANS attains 96% of queue stability region compared to the optimal exhaustive algorithm.

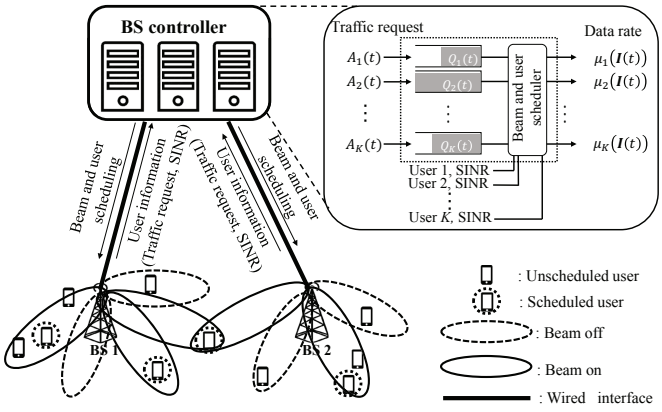


Fig. 1: Joint beam activation and user scheduling systems.

II. System Model

Network model. We consider a multi-cell downlink MIMO CoMP system, i.e., a user can be served by several base stations (BSs) controlled by a BS controller as shown in Fig. 1[9]. There are K users and N BSs controlled by a BS controller, and the set of all users and all BSs are denoted by $\mathcal{K} \doteq \{1, ..., K\}$ and $\mathcal{N} \doteq \{1, ..., N\}$, respectively. We assume a time-slotted TDMA (Time Division Multiple Access) system indexed by $t = \{0, 1, ...\}$ where the length of a time slot is Δt (in msec). During the slot, the channels for all links are assumed to be invariant. Each BS n has M antennas and a set of randomly generated beamforming vectors $\boldsymbol{b}$ from BS n is denoted by $\mathcal{B}_n \subseteq \mathcal{B}$.

Resource and link model. We assume that all beamforming vectors in BS n have equal fixed transmit power $p_{n_{\boldsymbol{b}}} = p_n =$

[5]Although the random beamforming does not optimally generate precoding vectors, it is known that as the number of antenna increases, the rate performance of random beamforming becomes asymptotically same as that of the optimal algorithm such as dirty paper coding (DPC) [15].

[6]For instance, we can predict average traffic arrivals at the same time duration of another day.

[7]Average queue backlog equals average arrival rate multiplied by average delay according to the Little's law [17].

[8]Notice that there are many studies to derive energy and queue bounds in Lyapunov framework for short-term problems of which optimal solution can be easily found within polynomial time (see e.g., [19] and references therein).

[9]This Cloud-RAN (Radio Access Network) style of network architecture is well accepted in 5G standards and industry [13].

$P_n/|\mathcal{B}_n|$ where P_n denotes total transmit power of BS n. Moreover, we denote $\mathcal{B}_k \subseteq \mathcal{B}$ by a set of beamforming vectors selected by user k, and $\mathcal{K}_n \subseteq \mathcal{K}$ by a set of users served by BS n. Each beamforming vector $\boldsymbol{b}$ possibly chooses at most one user, and one user can be allocated by multiple beamforming vectors at each time slot t. A set of pairs of a beamforming vector and a user is denoted by $\mathcal{I}(t)$. A scheduling indication function is denoted by $I_{\boldsymbol{b},k}(t)$ for a pair of beamforming vector $\boldsymbol{b}$ and user k, i.e., $I_{\boldsymbol{b},k}(t) = 1$ if beamforming vector $\boldsymbol{b}$ is activated and allocated to user k at time slot t, otherwise, $I_{\boldsymbol{b},k}(t) = 0$. A set of pairs of beamforming vector and user scheduling decisions is denoted by $\boldsymbol{I}(t) \doteq \{(\boldsymbol{b}, k)|I_{\boldsymbol{b},k}(t) = 1\}$, i.e., $\boldsymbol{I}(t) \subseteq \mathcal{I}(t)$.

We assume each user has one receive antenna without loss of generality, i.e., at most one data stream transmits to one user, then signal to interference plus noise ratio (SINR) γ_k of user k is denoted by

$$\gamma_k(\boldsymbol{I}(t)) = \frac{\sum_{\boldsymbol{b}\in\mathcal{B}_k} |\sqrt{p_{n_{\boldsymbol{b}}}}\boldsymbol{h}_{n_{\boldsymbol{b}},k}\boldsymbol{b}|^2 I_{\boldsymbol{b},k}(t)}{\phi_k + z_k^2}, \tag{1}$$

where $\boldsymbol{h}_{n_{\boldsymbol{b}},k} \in \boldsymbol{h}$ is a channel vector between user k and BS $n_{\boldsymbol{b}}$ which has beamforming vector $\boldsymbol{b}$. The channel vector is given by $\boldsymbol{h}_{n_{\boldsymbol{b}},k} = [\alpha_{n_{\boldsymbol{b}},k,1}, \alpha_{n_{\boldsymbol{b}},k,2}, \ldots, \alpha_{n_{\boldsymbol{b}},k,M}]\sqrt{\rho_{n_{\boldsymbol{b}},k}}$ where $\alpha_{n_{\boldsymbol{b}},k,m}$ is the multi-antenna channel co-efficient corresponding to BS $n_{\boldsymbol{b}}$, user k and antenna m, and $\rho_{n_{\boldsymbol{b}},k}$ is a large scale fading including path loss and shadowing, between BS $n_{\boldsymbol{b}}$ and user k. Denote by z_k^2 the thermal noise power of user k. Similar to [6] and [10], ϕ_k is the worst case interference of user k which is independent to the scheduling of other beams to make the formulation tractable. The worst case interference model works well on relatively low SNR systems because the influence of interference is much smaller compared to thermal noise. The achievable data rate of user k is given by Shannon's capacity formula [20] as follows.

$$\mu_k(\boldsymbol{I}(t)) = W \log_2(1 + \gamma_k(\boldsymbol{I}(t))), \tag{2}$$

where W denotes the entire system bandwidth.

Queue and energy model. Denote by $Q_k(t)$ the queue backlog of user k at time slot t. At the beginning of time slot t, $A_k(t)$ amount of traffic is arrived for user k. We should be noted that all users in the network generate different amount of $A_k(t)$ every time slot due to the traffic dynamics. Then, the queue backlogs for all users are updated as follows.

$$Q_k(t+1) = \left[Q_k(t) - \mu_k(\boldsymbol{I}(t)) + A_k(t)\right]^+ \text{[bits]}, \ \ \forall k \in \mathcal{K}, \tag{3}$$

where $[\cdot]^+$ denotes the projection onto the set of non-negative real numbers and $\lambda_k = \mathbb{E}\{A_k(t)\}$ denotes average arrival rate of user k and $\lambda^{in} = \{\mathbb{E}\{A_1(t)\}, \mathbb{E}\{A_2(t)\}, ..., \mathbb{E}\{A_K(t)\}\}$ denotes average arrival rate vector for all users.

According to BS energy consumption model in [21], the transmit power exerts substantial influence on the required power for amplifier, cooling systems, and so on, where the influence is often linear; hence we take account of only transmit power allocated to each beam in this paper. Then, we define total energy consumption with respect to transmit power during one time slot as follows.

$$E(\boldsymbol{I}(t)) = \sum_{k\in\mathcal{K}} \sum_{\boldsymbol{b}\in\mathcal{B}_k} p_{n_{\boldsymbol{b}}} I_{\boldsymbol{b},k}(t)\Delta t. \tag{4}$$

In the beginning of every time slot, every user sends their SINR and traffic request (or arrival) information to centralized BS controller via the one of associated BSs (see Fig. 1). The centralized BS controller decides beamforming activation and user scheduling by exploiting the SINR, traffic arrival and queue backlogs of all users. The decided beam activation and user scheduling outputs are informed from the centralized BS controller to each BS.

III. Joint Beam and User Scheduling Algorithm

In this section, we formulate an optimization problem considering energy minimization while ensuring finite service time with respect to dynamic traffic arrivals. Then, we develop a joint beam on/off activation and user scheduling, namely BEANS algorithm.

A. Problem Formulation

Our objective in a multi-cell cooperative network is to minimize average energy consumption while stabilizing queue backlogs of all users for all arrival rate vectors within capacity region. The capacity region is defined as a set of all arrival rate vectors for which there exists certain control scheme which is able to support the traffic arrivals within finite service time. We formally state an optimization problem as follows.

$$\textbf{(P):} \quad \min_{\boldsymbol{I}=(\boldsymbol{I}(t))_{t=0}^{T-1}} \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{E(\boldsymbol{I}(t))\}, \tag{5}$$

$$s.t. \quad \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{k\in\mathcal{K}} \mathbb{E}\{Q_k(t)\} < \infty, \tag{6}$$

$$\sum_{k\in\mathcal{K}_{\boldsymbol{b}}} I_{\boldsymbol{b},k}(t) \le 1, \ \forall t, \forall \boldsymbol{b} \in \mathcal{B}, \tag{7}$$

$$\sum_{k\in\mathcal{K}_n} \sum_{\boldsymbol{b}\in\mathcal{B}_n} I_{\boldsymbol{b},k}(t) \ge 1, \ \forall t, \forall n \in \mathcal{N}, \tag{8}$$

where $\mathcal{K}_{\boldsymbol{b}}$ is a set of all users which can be served by beamforming vector $\boldsymbol{b}$. Constraint (6) means the system should ensure a finite service time with respect to dynamic traffic arrivals and normally, if the system meets this constraint, we say that queues are stable [16]. Constraints (7) and (8) mean each beamforming vector can be allocated to at most one user and each BS should schedule at least one user, respectively.

B. Algorithm Design

To convert our long-term problem **(P)** into online-fashioned slot-by-slot problem, we invoke Lyapunov optimization technique [16]. We first define Lyapunov function and Lyapunov drift function to capture queue stability in equation (6) as follows.

$$L(t) \triangleq \tfrac{1}{2} \textstyle\sum_{k\in\mathcal{K}} [Q_k(t)]^2, \tag{9}$$

$$\Delta(L(t)) \triangleq \mathbb{E}\{L(t+1) - L(t)|\mathbf{Q}(t)\}. \tag{10}$$

To capture the minimization of average energy consumption in equation (5), we adopt drift-plus-penalty approach in [16] where the penalty function is the sum of expected energy consumption of all BSs as follows.

$$\Delta(L(t)) + V\mathbb{E}\{E(\boldsymbol{I}(t))\}, \tag{11}$$

where V is an energy-delay tradeoff parameter. We derive upper bound of (11) which is a standard approach of Lyapunov optimization by the following Lemma.

Lemma 1. Under any possible control variables $\boldsymbol{I}(t)$, we have:

$$\Delta(L(t)) + V\mathbb{E}\{E(t)|\mathbf{Q}(t)\} \leq B + \sum_{k\in\mathcal{K}} Q_k(t)\lambda_k - \mathbb{E}\Big\{\sum_{k\in\mathcal{K}} Q_k(t)\mu_k(\boldsymbol{I}(t))|\mathbf{Q}(t)\Big\} + V\mathbb{E}\{E(\boldsymbol{I}(t))|\mathbf{Q}(t)\}, \quad (12)$$

where $B = \frac{K}{2}\big(A_{max}^2 + \mu_{max}^2\big)$, and A_{max} and μ_{max} denote maximum traffic arrival and maximum data rate of a user at a time slot, respectively.

Proof: It can be easily proven using queueing dynamics in equation (3), e.g., see [19]. ■

By minimizing RHS of (12) every time slot, we can realize the slot-by-slot optimization of original problem (**P**) for given energy-delay tradeoff parameter V. Then, slot-by-slot problem (**SBSP**) is given by

(SBSP):

$$\max_{\boldsymbol{I}(t)\in\mathcal{I}(t)} G(\boldsymbol{I}(t)) = \mathbb{E}\Big\{\sum_{k\in\mathcal{K}} Q_k(t)\mu_k(\boldsymbol{I}(t))\Big\} - V\mathbb{E}\{E(\boldsymbol{I}(t))\},$$
$$s.t. \quad \text{constraints (7), (8),}$$

where $\mathcal{I}(t) \doteq \{(\boldsymbol{b},k)|\boldsymbol{b}\in\mathcal{B}_k, k\in\mathcal{K}\}$. Although we derive **(SBSP)** by the time slot-based manner, these modified max-weight style problems are known as NP-hard [18]. However, the work in [22] provided a constant factor approximation algorithm with polynomial time complexity aiming to maximize non-negative submodular function under matroids constraint. The definition of the submodularity is as follows.

Definition 1 (Submodularity)**.** Let $\mathcal{G}$ be a finite ground set and $h: 2^{\mathcal{G}} \to \mathbb{R}$. Then h is submodular if for all $\mathcal{A}, \mathcal{B} \subseteq \mathcal{G}$,

$$h(\mathcal{A}) + h(\mathcal{B}) \geq h(\mathcal{A}\cup\mathcal{B}) + h(\mathcal{A}\cap\mathcal{B}),$$

or equivalently, for every $\mathcal{A}\subseteq\mathcal{G}$ and $x_1, x_2 \in \mathcal{G}\setminus\mathcal{A}$,

$$h(\mathcal{A}\cup\{x_1\}) + h(\mathcal{A}\cup\{x_2\}) \geq h(\mathcal{A}\cup\{x_1,x_2\}) + h(\mathcal{A}).$$

We can easily know our objective function $G(\cdot)$ is a submodular function as well by the Definition 1.

C. Joint Beam Activation and User Scheduling Algorithm

Using the fact that the function $G(\cdot)$ is submodular and based on the work in [22], we provide a greedy-style joint beam activation and user scheduling, namely BEANS algorithm. This jointly controls beam on/off activation and user scheduling every time slot t without knowledge of future network states such as traffic arrivals. To describe BEANS algorithm, we define a set of feasible beam and user scheduling pairs $\mathcal{Y}(t) \doteq \{\boldsymbol{I}(t)|(\boldsymbol{b},k)\in\boldsymbol{I}(t), \sum_{k\in\mathcal{K}} I_{\boldsymbol{b},k}(t) \leq 1\}$, i.e., $\mathcal{Y}(t)$ is the set of all $\boldsymbol{I}(t)$ satisfying constraint (7). The BEANS algorithm can be represented as follows.

BEANS (Joint BEam Activation aNd user Scheduling)

Every time slot t,

Initialization. Let $f(\boldsymbol{I}(t)) \triangleq G(\boldsymbol{I}(t))$ and set $\mathcal{I}(t)_1 := \mathcal{I}(t)$.

Step 1: For $j = 1, 2$, do:

(a) Apply the **LSP** on the ground set $\mathcal{I}(t)_j$ and function f to obtain a solution $\boldsymbol{I}_j(t) \subseteq \mathcal{I}(t)_j$ corresponding to the problem:

$$\boldsymbol{I}_j(t): \max_{\boldsymbol{I}(t)\subseteq\mathcal{I}(t)_j} \{f(\boldsymbol{I}(t))|\boldsymbol{I}(t)\in\mathcal{Y}(t)\}$$

(b) Set $\mathcal{I}(t)_{j+1} := \mathcal{I}(t)_j \setminus \boldsymbol{I}_j(t)$.

(c) Define $\mathcal{N}'$ as a set of BSs which do not have any element of $\boldsymbol{I}_j(t)$'s beamforming vector. If $\mathcal{N}' \neq \{\varnothing\}$, then set $s := \max_{(\boldsymbol{b},k)\in\mathcal{I}(t)_j} f(\boldsymbol{I}_j(t)\cup(\boldsymbol{b},k))$, s.t. $n_{\boldsymbol{b}} \in \mathcal{N}'$, update $\boldsymbol{I}_j(t) := \boldsymbol{I}_j(t)\cup\{s\}$, and go back to (c), otherwise, go to **Step 2**.

Step 2: Return the solution corresponding to

$$\boldsymbol{I}_{BEANS}(t) := \arg\max\{f(\boldsymbol{I}_1(t)), f(\boldsymbol{I}_2(t))\}.$$

Local Search Procedure (LSP)

Input: Ground set $\mathcal{X}$ of elements and function f.

Initialization. Set $\boldsymbol{I}(t) := \max_{(\boldsymbol{b},k)\in\mathcal{X}} f((\boldsymbol{b},k))$.

Step 1: Set $s := \max_{(\boldsymbol{b},k)\in\mathcal{X}} f((\boldsymbol{I}(t)\setminus(\boldsymbol{b'},k'))\cup(\boldsymbol{b},k))$, s.t. $(\boldsymbol{I}(t)\setminus(\boldsymbol{b'},k'))\cup(\boldsymbol{b},k)\in\mathcal{Y}(t), (\boldsymbol{b'},k')\in\boldsymbol{I}(t)\cup\{\varnothing\}$. If $f((\boldsymbol{I}(t)\setminus(\boldsymbol{b'},k'))\cup\{s\}) > (1+\frac{\varepsilon}{|\mathcal{I}(t)|^4})f(\boldsymbol{I}(t))$, then update $\boldsymbol{I}(t) := (\boldsymbol{I}(t)\setminus(\boldsymbol{b'},k'))\cup\{s\}$ and go back to **Step 1**, otherwise, go to **Step 2**.

Step 2: Set $s := \max_{(\boldsymbol{b},k)\in\mathcal{X}} f(\boldsymbol{I}(t)\setminus(\boldsymbol{b},k))$. If $f(\boldsymbol{I}(t)\setminus\{s\}) > (1+\frac{\varepsilon}{|\mathcal{I}(t)|^4})f(\boldsymbol{I}(t))$ then update $\boldsymbol{I}(t) := \boldsymbol{I}(t)\cup\{s\}$ and go back to **Step 1**, otherwise, return $\boldsymbol{I}(t)$.

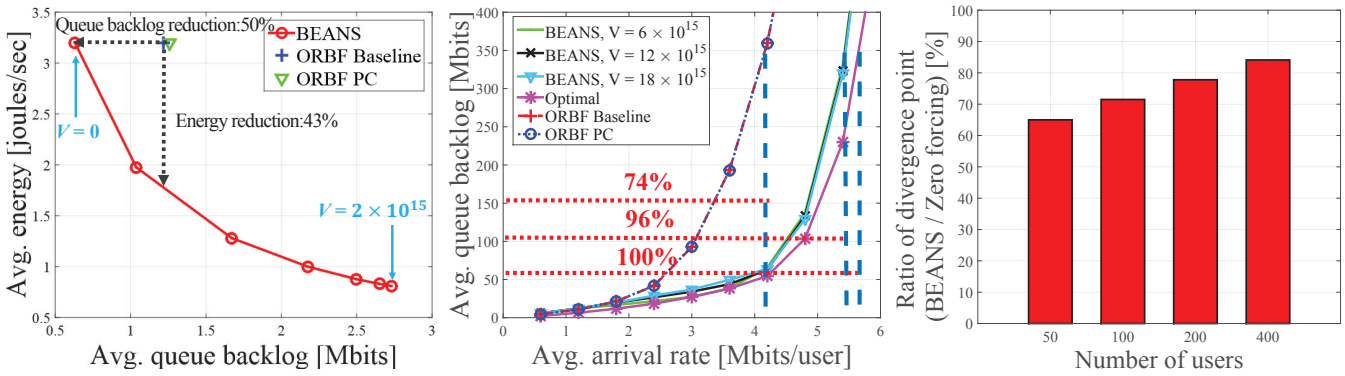
where ε denotes a very small constant to guarantee polynomial time complexity of BEANS algorithm. **Local Search Procedure (LSP)** adds the elements one by one by a greedy manner under the condition of which $\boldsymbol{I}(t)$ satisfies constraint (7). If $f(\boldsymbol{I}(t))$ cannot be larger anymore by adding the some element, the **LSP** will search whether there is increment of $f(\boldsymbol{I}(t))$ by subtracting the existing element or not. At the process (c) of **Step 1** in BEANS, if there is a BS which does not schedule any user, $\boldsymbol{I}_j(t)$ adds elements one by one by a greedy manner under the condition of which $\boldsymbol{I}_j(t)$ satisfies constraint (8), otherwise, this algorithm ends by returning $\boldsymbol{I}_{BEANS}(t)$.

However, unfortunately, it is known that there is no polynomial time algorithm which attains constant factor approximation to the optimal performance for a submodular function *with negative term*, which is the same form as our function $G(\boldsymbol{I}(t))$ in slot-by-slot objective **(SBSP)** [23]. Hence, in the next subsection, we demonstrate performance bounds of BEANS algorithm in terms of long-term average queue backlog and energy consumption by utilizing alternative asymptotic approach.

D. Theoretical Analysis

We clearly show the set of arrival rate vectors which can be supported by BEANS in Theorem 1.

Theorem 1. BEANS guarantees queue stability for all users if an arrival rate vector is within $\frac{1}{2+\varepsilon}$ of capacity region.



(a) Energy-delay tradeoff. (b) Queue stability region. (c) Zero forcing vs. random beamforming.

Fig. 2: Simulation results for all scenarios: 2D topology for (a), linear topologies for (b) and (c).

Proof: The proof is presented in the Appendix A. ■

Although the short-term performance bounds of BEANS cannot be shown in general [23], we demonstrate that BEANS algorithm attains constant-ratio approximation of capacity region by using the feature of queueing dynamics in Theorem 1. In addition, Theorem 2 provides average energy consumption and queue backlog performance of BEANS.

Theorem 2. Whenever arrival rate vector λ^{in} is within $\frac{1}{4+\varepsilon}$ of capacity region, under BEANS algorithm, we have:
(average energy consumption):

$$\limsup_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\{E(t)\}\leq$$
$$\frac{B}{V}+\frac{3+\varepsilon}{4+\varepsilon}\sum_{n\in\mathcal{N}}P_n\Delta t+\sum_{n\in\mathcal{N}}p_n\Delta t+\frac{1}{4+\varepsilon}\psi(\lambda^{out}+\sigma(\lambda^{out})),$$

(average queue backlog):

$$\limsup_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T-1}\sum_{k\in\mathcal{K}}\mathbb{E}\{Q_k(t)\}\leq\frac{4+\varepsilon}{\sigma(\lambda^{out})}B+\frac{3+\varepsilon}{\sigma(\lambda^{out})}V\sum_{n\in\mathcal{N}}P_n\Delta t$$
$$+\frac{4+\varepsilon}{\sigma(\lambda^{out})}V\sum_{n\in\mathcal{N}}p_n\Delta t+\frac{1}{\sigma(\lambda^{out})}V\psi(\lambda^{out}+\sigma(\lambda^{out})),$$

where $\psi(\lambda^{out}+\sigma(\lambda^{out}))$ is a minimumum energy consumption with queue stability when traffic arrivals vector is $\lambda^{out}=(4+\varepsilon)\lambda^{in}$.

Proof: The proof is presented in the Appendix B. ■

The result of Theorem 2 shows that as the parameter V increases, the average energy consumption decreases whereas the average queue backlogs increases, vice versa.

IV. Performance Evaluation

In this section, we execute extensive simulations to demonstrate the performance of the BEANS algorithm.

Setup. We consider two topologies: *2D topology*, where 4 BSs (each of which has 4 antennas) and 100 users are randomly located in the 400m×400m with uniform distribution, and *linear topology*, where 2 BSs (each of which has 2 antennas) and various number of users (depending on the scenario) are randomly located in the 400m line with uniform distribution.

For all scenarios, each BS has 23 dBm transmit power budget and the system bandwidth is 10 MHz. The path loss is set to be $128.1+37.6\log_{10}(d)$ where d is the distance from BS to user in [km] and the standard deviation of shadowing is set to be 8 dB. Each channel coefficient, $[\alpha_{n_b,k,1},\alpha_{n_b,k,2},\ldots,\alpha_{n_b,k,M}]$, is modeled as a zero mean complex Gaussian random variable. We adopt a simple clustering technique for the cooperative networks: a user can be associated to any BSs if the distance between the user and the BSs are less than the threshold: $100\sqrt{2}$m.

We compare BEANS algorithm with existing algorithms [14], [24] and unrealistic optimal algorithm. In ORBF baseline algorithm [14], each user is matched to each beam which attains highest SINR, and the users are randomly scheduled every time slot. In ORBF PC algorithm [24], the algorithm procedure is the same as ORBF baseline except for power control. In zero forcing algorithm, each user is matched to each beam corresponding to zero forcing beamforming scheme [25], and the user scheduling follows simple greedy algorithm. Optimal algorithm exhaustively searches all beam and user scheduling combinations which satisfy **(SBSP)**.

We consider average energy consumption and average queue backlog as two metrics in the simulations. For the first simulation, we show the energy-delay tradeoff which can be changed depending on the tradeoff parameter V. We should be noted that average queue backlog is indirectly interpreted as average delay according to the Little's law [17]. For the second and third simulations, we demonstrate the performance of queue stability (i.e., divergence point of queue backlog) with respect to the average arrival rates, and this performance implies that how much arrival rates can be stably served by the system within capacity region.

Simulation results. We present our results by summarizing the key observations.

Energy-delay tradeoff. Fig. 2(a) shows the tradeoff between average queue backlog and energy consumption. BEANS algorithm can save 43% of energy consumption for the same average queue backlog, and reduce 50% of average queue backlog for the same energy consumption compared to ORBF baseline and ORBF PC algorithms. This is due to the fact that there are no knowledge of energy-delay tradeoff and

queue backlog in ORBF baseline and ORBF PC algorithms. In addition, BEANS algorithm can significantly save energy consumption when average arrival rate is far from capacity boundary, i.e., low average arrival rate.

Queue stability. Fig. 2(b) depicts average queue backlog versus average arrival rates. Notice that divergence point, i.e., average arrival rate when the queue backlog starts to diverge[10], of the optimal algorithm means capacity region of this system. The average queue backlog of BEANS diverges at 96% of capacity region which implies BEANS algorithm shows near optimal performance compared to the optimal algorithm whereas the performances of other algorithms such as ORBF baseline and ORBF PC are far from the optimal performance. In addition, BEANS attains similar divergence points for different Vs which means higher V does not deteriorate queue stability of all users.

Zero forcing versus orthogonal random beamforming. Fig. 2(c) shows a ratio of the BEANS divergence point over zero-forcing divergence point as a function of number of users. As the number of users increases, the performance of BEANS algorithm gets closer to that of zero forcing algorithm. This implies that simple BEANS algorithm attains the similar performance with zero forcing algorithm under the high user density scenario even though the zero forcing algorithm is difficult to be implemented due to full CSI requirements from all users.

V. Conclusion

As a way to save energy expenditure of base stations, we propose a joint beam and user scheduling policy, namely BEANS in cooperative cellular network systems. Proposed policy makes an effort to minimize average energy consumption of base stations while ensuring long-term delay performance of all users by means of an adaptive operation corresponding to unpredictably changing traffic arrivals and channel states. We believe that smart network management such as proposed beam/user scheduling policy would have a great opportunity to save energy consumption of network operation in multiple antenna cooperative cellular networks which will be a main part of future 5G network systems.

VI. Appendix

A. Proof of Theorem 1.

Proof: Suppose an arrival rate vector, λ^{in}, is located within $\frac{1}{2+\varepsilon}$ of capacity region and there is a set of users $\mathcal{U} \subseteq \mathcal{K}$ whose queue backlogs diverge as time goes by and the queue backlogs of users who belong to $\mathcal{K} \setminus \mathcal{U}$ do not diverge. Then, there exists $\alpha < \infty$ such that all users who belong to $\mathcal{K} \setminus \mathcal{U}$ satisfying the following inequality.

$$Q_k(t) \leq \alpha,\ w.p.1,\ \forall k \in \mathcal{K} \setminus \mathcal{U}, \tag{13}$$

where $w.p.1$ means with probability 1.

[10]Indeed, the average data rate where the average queue backlog is abruptly increasing is the divergence point, but here, the average arrival rate when the average queue backlog exceeds 350 Mbits is designated as the divergence point for simplicity.

Lemma 2. If the above assumption is correct, there exist $\beta < \infty$ and positive integer $T < \infty$ satisfying the following inequalities when we assume there are lower and upper bounds of achievable increasing data rate for all users, i.e., $\mu_{d,min} < \mu_k(\boldsymbol{I}(t) \cup (\boldsymbol{b}, k)) - \mu_k(\boldsymbol{I}(t)) < \mu_{d,max},\ \forall (\boldsymbol{b}, k) \notin \boldsymbol{I}(t),\ \boldsymbol{I}(t) \cup (\boldsymbol{b}, k) \subseteq \mathcal{I}(t)$.

$$Q_k(T) < \infty,\ \forall k \in \mathcal{U}, \tag{14}$$

$$Q_k(t) \geq \beta,\ \forall k \in \mathcal{U},\ \forall t \geq T, \tag{15}$$

$$\beta \mu_{d,min} > \alpha \mu_{d,max}, \tag{16}$$

$$\beta \mu_{d,min} > V p_n \Delta t,\ \forall n \in \mathcal{N}. \tag{17}$$

Proof: Because all $k \in \mathcal{U}$ have queue backlogs which diverge as time goes by, if there is no T which satisfies (14) and (15) for given β which satisfies (16) and (17), this contradicts to the first assumption. This completes the proof. ■

Denote by $\mathcal{B}_{\mathcal{U}}$ all beamforming vectors which can be allocated to a set of users $\mathcal{U}$. Since BEANS is operated by a greedy manner, when t is greater or equal to T, we first assign all beams which can be matched with $\mathcal{U}$ to $\mathcal{B}_{\mathcal{U}}$, and then assign the remaining beams which can be matched with $\mathcal{K} \setminus \mathcal{U}$ to $\mathcal{B}_{\mathcal{U}}$ in the BEANS algorithm. This is because (16) means any beam allocation on $\mathcal{U}$ is always better than that on $\mathcal{K} \setminus \mathcal{U}$. Therefore, we first focus on beam scheduling for $\mathcal{U}$ to derive the performance of BEANS. We formulate a problem **(SBSP-D)** as follows.

(SBSP-D):

$$\max_{\boldsymbol{I}(t) \in \mathcal{I}(t)} \Big(G(\boldsymbol{I}(t))_{\boldsymbol{D}} = \mathbb{E}\Big\{ \sum_{k \in \mathcal{U}} Q_k(t) \mu_k(\boldsymbol{I}(t)) \Big\} - V \mathbb{E}\{E(\boldsymbol{I}(t))\} \Big),$$

$$s.t. \quad \sum_{k \in \mathcal{U}_{\boldsymbol{b}}} I_{\boldsymbol{b},k}(t) \leq 1,\ \forall \boldsymbol{b} \in \mathcal{B}_{\mathcal{U}}, \tag{18}$$

$$\sum_{k \in \mathcal{U}_n} \sum_{\boldsymbol{b} \in \mathcal{B}_n} I_{\boldsymbol{b},k}(t) \geq 1,\ \forall n \in \mathcal{N}_{\mathcal{U}}, \tag{19}$$

where $\mathcal{N}_{\mathcal{U}}$ denotes the set of all BSs which can serve a set of users $\mathcal{U}$ for $t \geq T$. Now, we introduce a modified algorithm, namely BEANS-D, for ease of explanation.

BEANS-D

Every time slot t,

Initialization. Let $f(\boldsymbol{I}(t)) \triangleq G(\boldsymbol{I}(t))_{\boldsymbol{D}}$ and set $\mathcal{I}(t)_1 := \mathcal{I}(t)$.

Step 1: For $j = 1, 2$, do:

(a) Apply the **LSP** on the ground set $\mathcal{I}(t)_j$ and function f to obtain a solution $\boldsymbol{I}_j(t) \subseteq \mathcal{I}(t)_j$ corresponding to the problem:

$$\boldsymbol{I}_j(t) : \max_{\boldsymbol{I}(t) \subseteq \mathcal{I}(t)_j} \{ f(\boldsymbol{I}(t)) | \boldsymbol{I}(t) \in \mathcal{Y}(t) \}$$

(b) Set $\mathcal{I}(t)_{j+1} := \mathcal{I}(t)_j \setminus \boldsymbol{I}_j(t)$.

Step 2: Return the solution corresponding to

$$\boldsymbol{I}_{\mathbf{D}}(t) := \arg\max \{ f(\boldsymbol{I}_1(t)), f(\boldsymbol{I}_2(t)) \}.$$

The difference between BEANS-D and BEANS is that the process (c) of **Step 1** in BEANS, which is to satisfy the

constraint (19), is eliminated in BEANS-D. In other words, BEANS-D is an algorithm to solve the **(SBSP-D)** problem except for the constraint (19). If the **(SBSP-D)** problem does not have constraint (19), this problem is the same as maximizing nonnegative monotone submodular function under matroid constraint. This is because that constraint (18) first produces a matroid constraint, and $G(\boldsymbol{I}(t))_{\boldsymbol{D}}$ always has a nonnegative value due to the condition (17). Then, by invoking the result of [22], BEANS-D satisfies the following inequality for $G(\boldsymbol{I}(t))_{\boldsymbol{D}}$.

$$G(\boldsymbol{I_D}(t))_{\boldsymbol{D}} \geq \frac{1}{2+\varepsilon} G(\boldsymbol{I}^*_{\boldsymbol{D}}(t))_{\boldsymbol{D}},$$

where $\boldsymbol{I}^*_{\boldsymbol{D}}(t)$ denotes optimal solution of **(SBSP-D)** problem except for the constraint (19). Also, we can check that the following equalities.

$$\boldsymbol{I}^{\mathcal{U}}_{BEANS}(t) = \boldsymbol{I_D}(t), \quad \boldsymbol{I}^*(t) = \boldsymbol{I}^*_{\boldsymbol{D}}(t),$$

where $\boldsymbol{I}^{\mathcal{U}}_{BEANS}(t) \doteq \{(\boldsymbol{b},k)|(\boldsymbol{b},k) \in \boldsymbol{I}_{BEANS}(t), k \in \mathcal{U}\}$ and $\boldsymbol{I}^*(t)$ means the optimal solution of **(SBSP-D)** problem. As mentioned before, $\boldsymbol{I_D}(t)$ is the solution of **(SBSP-D)** problem except for the constraint (19) and $\boldsymbol{I}_{BEANS}(t)$ is the solution with a consideration of constraint (19). However, BEANS and BEANS-D are algorithms that operate in a greedy manner, hence each BS tries to select as many users as possible due to the condition (17). Therefore, even if the constraint (19) is not created, the result of the above algorithm, $\boldsymbol{I_D}(t)$, is presented in the set of solutions satisfying the constraint (19). The relation between $\boldsymbol{I}^*(t)$ and $\boldsymbol{I}^*_{\boldsymbol{D}}(t)$ can be explained in a similar way. Therefore, we can derive the following results.

$$G(\boldsymbol{I}^{\mathcal{U}}_{BEANS}(t))_{\boldsymbol{D}} \geq \frac{1}{2+\varepsilon} G(\boldsymbol{I}^*(t))_{\boldsymbol{D}}.$$

We can use (12) to express the following inequality about $\mathcal{U}$ using the fact that BEANS guarantees $\frac{1}{2+\varepsilon}$-optimal and then,

$$\begin{aligned}
&\Delta(L(t)) + V\mathbb{E}\{E(t)|\mathbf{Q}(t)\} \\
&\leq B + \sum_{k\in\mathcal{U}} Q_k(t)\lambda^{in}_k - G(\boldsymbol{I}(t)^*)_{\boldsymbol{D}} \\
&\leq B + \sum_{k\in\mathcal{U}} Q_k(t)\lambda^{in}_k - G(\boldsymbol{I}^{\mathcal{U}}_{BEANS}(t))_{\boldsymbol{D}} \qquad (20)\\
&\leq B + \sum_{k\in\mathcal{U}} Q_k(t)\lambda^{in}_k - \frac{1}{2+\varepsilon} G(\boldsymbol{I}(t)^*)_{\boldsymbol{D}}.
\end{aligned}$$

Moreover, we introduce S-only algorithm which is known to minimize energy while achieving queue stability for all arrival rate vectors within a capacity region [16]. Then, we have:

$$\begin{aligned}
&B + \sum_{k\in\mathcal{U}} Q_k(t)\lambda^{in}_k - \frac{1}{2+\varepsilon} G(\boldsymbol{I}(t)^*)_{\boldsymbol{D}} \\
&\leq B + \sum_{k\in\mathcal{U}} Q_k(t)\lambda^{in}_k + \frac{1}{2+\varepsilon}[V\mathbb{E}\{E(\boldsymbol{I}(t)_{S-only})\} \\
&- \mathbb{E}\{\sum_{k\in\mathcal{U}} Q_k(t)\mu_k(\boldsymbol{I}(t)_{S-only})|\mathbf{Q}(t)\}] \qquad (21)\\
&= B + \frac{1}{2+\varepsilon}[V\psi(\lambda^{out} + \sigma(\lambda^{out})) \\
&- \sum_{k\in\mathcal{U}} Q_k(t)[\sigma(\lambda^{out}) + \lambda^{out}_k - (2+\varepsilon)\lambda^{in}_k],
\end{aligned}$$

where $\psi(\lambda^{out} + \sigma(\lambda^{out}))$ denotes minimum BSs' energy consumption when arrival rate vector is λ^{out} and $\sigma(\lambda^{out})$ represents a measure of the distance between the rate vector λ^{out} and the capacity boundary [16]. We can derive this equation by using the similar derivation process with [16]. Then, we have:

$$\begin{aligned}
&\limsup_{T'\to\infty} \frac{1}{T'-T} \sum_{t=T}^{T'-1} \sum_{k\in\mathcal{U}} \mathbb{E}\{Q_k(t)\}[\sigma(\lambda^{out}) + \lambda^{out}_k \qquad (22)\\
&- (2+\varepsilon)\lambda^{in}_k] \leq (2+\varepsilon)B + V\cdot\psi(\lambda^{out} + \sigma(\lambda^{out})).
\end{aligned}$$

Therefore BEANS sufficiently achieves queue stability if $\sigma(\lambda^{out}) + \lambda^{out}_k - (2+\varepsilon)\lambda^{in}_k > 0$. Notice that $\sigma(\lambda^{out})$ is always positive and λ^{out} can be any service vector if $(\lambda^{out} + \sigma(\lambda^{out}))$ is interior to capacity region. Thus, if traffic arrival vector λ^{in} satisfies following relation (23) with respect to any service vector λ^{out} interior to capacity region, each queue $Q_k(t),\ \forall k \in \mathcal{U}$ does not diverge as time goes by. It is a contradiction to the assumption in the beginning of the proof.

$$\lambda^{out}_k = (2+\varepsilon)\lambda^{in}_k,\ \forall k \in \mathcal{U}. \qquad (23)$$

This completes the proof. ■

B. Proof of Theorem 2.

Proof: Since $G(\boldsymbol{I}(t))$ has a negative term, there is no known polynomial time algorithm which guarantees constant lower bound of an optimal performance [23]. Hence, we can think of a way to add a positive term to $G(\boldsymbol{I}(t))$ to resolve this problem. We consider the following alternative problem (SBSP-α).

$$\begin{aligned}
\textbf{(SBSP-}\alpha\textbf{):} \quad &\max_{\boldsymbol{I}(t)\in\mathcal{I}(t)} \quad G(\boldsymbol{I}(t))_\alpha, \\
&s.t. \quad \text{constraint } (7),
\end{aligned}$$

$$G(\boldsymbol{I}(t))_\alpha = \mathbb{E}\Big\{\sum_{k\in\mathcal{K}} Q_k(t)\mu_k(\boldsymbol{I}(t))\Big\} - V\mathbb{E}\{E(\boldsymbol{I}(t))\} + V\sum_{n\in\mathcal{N}} P_n\Delta t.$$

We can see that the above **(SBSP-α)** is the same as maximizing nonnegative submodular function under the matroid constraint, and $G(\boldsymbol{I}(t))_\alpha$ is always greater than or equal to 0 because we added a positive constant value $V\sum_{n\in\mathcal{N}} P_n\Delta t$ which is always greater than or equal to $V\mathbb{E}\{E(\boldsymbol{I}(t))\}$. We can solve **(SBSP-$\alpha$)** by adding $G(\boldsymbol{I}(t))_\alpha$ instead of $G(\boldsymbol{I}(t))_{\boldsymbol{D}}$ in the **Initialization** of the BEANS-D algorithm. We call this algorithm as BEANS-α and the solution scheduling set is denoted by $\boldsymbol{I}_\alpha(t)$. Then, by invoking the result of [22], BEANS-α satisfies the following inequality for $G(\boldsymbol{I}(t))_\alpha$.

$$G(\boldsymbol{I}_\alpha(t))_\alpha \geq \frac{1}{4+\varepsilon} G(\boldsymbol{I}^*_\alpha(t))_\alpha,$$

where $\boldsymbol{I}^*_\alpha(t)$ means the optimal solution of **(SBSP-α)**. Using $G(\boldsymbol{I}(t)) = G(\boldsymbol{I}(t))_\alpha - V\sum_{n\in\mathcal{N}} P_n\Delta t$, we derive the following inequality.

$$G(\boldsymbol{I}_\alpha(t)) \geq \frac{1}{4+\varepsilon} G(\boldsymbol{I}^*_\alpha(t)) - \frac{3+\varepsilon}{4+\varepsilon} V\sum_{n\in\mathcal{N}} P_n\Delta t.$$

Moreover, we can present the relation between $\boldsymbol{I}_\alpha(t)$ and $\boldsymbol{I}_{BEANS}(t)$ as follows.

$$G(\boldsymbol{I}_{BEANS}(t)) \geq G(\boldsymbol{I}_\alpha(t)) - V\sum_{n\in\mathcal{N}} p_n\Delta t.$$

This is because BEANS has process (c) in **Step 1** which captures the constraint (8), whereas BEANS-α skips this process. Finally, using the fact that the $G(\boldsymbol{I}_{\alpha}^{*}(t))$ except for the constraint (8) is always larger than the $G(\boldsymbol{I}^{*}(t))$ which is the optimal value with the constraint (8), the following inequality is derived.

$$G(\boldsymbol{I}_{BEANS}(t)) \geq \frac{1}{4+\varepsilon} G(\boldsymbol{I}^{*}(t)) - \frac{3+\varepsilon}{4+\varepsilon} V \sum_{n\in\mathcal{N}} P_n \Delta t - V \sum_{n\in\mathcal{N}} p_n \Delta t. \tag{24}$$

Then, similar to (20), (21) and by invoking [16], we can derive following inequality.

$$\begin{aligned}
&\mathbb{E}\{L(\mathbf{Q}(T))\} - \mathbb{E}\{L(\mathbf{Q}(0))\} + V \sum_{t=0}^{T-1} \mathbb{E}\{E(t)\} \\
&\leq BT + \frac{VT}{4+\varepsilon}\psi(\lambda^{out} + \sigma(\lambda^{out})) + \frac{3+\varepsilon}{4+\varepsilon} VT \sum_{n\in\mathcal{N}} P_n \Delta t \\
&- \frac{1}{4+\varepsilon} \sum_{t=0}^{T-1} \sum_{k\in\mathcal{K}} \mathbb{E}\{Q_k(t)\}[\sigma(\lambda^{out}) + \lambda_k^{out} - (4+\varepsilon)\lambda_k^{in}] \\
&+ VT \sum_{n\in\mathcal{N}} p_n \Delta t.
\end{aligned}$$

If we assume $\lambda_k^{out} = (4+\varepsilon)\lambda_k^{in}$, we can derive upper bound of average energy consumption and the average queue backlog as follows.

(average energy consumption):

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\{E(t)\} \leq \frac{B}{V} + \frac{3+\varepsilon}{4+\varepsilon} \sum_{n\in\mathcal{N}} P_n \Delta t + \sum_{n\in\mathcal{N}} p_n \Delta t + \frac{1}{4+\varepsilon}\psi(\lambda^{out} + \sigma(\lambda^{out})),$$

(average queue backlog):

$$\limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \sum_{k\in\mathcal{K}} \mathbb{E}\{Q_k(t)\} \leq \frac{4+\varepsilon}{\sigma(\lambda^{out})} B + \frac{3+\varepsilon}{\sigma(\lambda^{out})} V \sum_{n\in\mathcal{N}} P_n \Delta t + \frac{4+\varepsilon}{\sigma(\lambda^{out})} V \sum_{n\in\mathcal{N}} p_n \Delta t + \frac{1}{\sigma(\lambda^{out})} V \psi(\lambda^{out} + \sigma(\lambda^{out})).$$

This completes the proof. ∎

ACKNOWLEDGEMENT

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (No.B0717-17-0034, Versatile Network System Architecture for Multidimensional Diversity). Also, this work was supported by the ICT R&D program of MSIP/IITP. [2016-0-00563, Research on Adaptive Machine Learning Technology Development for Intelligent Autonomous Digital Companion].

REFERENCES

[1] Cisco. San Jose, CA, "Cisco visual networking index: Global mobile data traffic forecast update, 2015-2020." [Online]. Available: http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html

[2] T. C. Group, "SMART 2020: Enabling the low carbon economy in the information age." [Online]. Available: http://www.smart2020.org/_assets/files/02_Smart2020Report.pdf

[3] H. Holtkamp, G. Auer, S. Bazzi, and H. Haas, "Minimizing base station power consumption," *IEEE JSAC*, vol. 32, no. 2, pp. 297–306, Feb. 2014.

[4] J. Kwak, K. Son, Y. Yi, and S. Chong, "Greening effect of spatio-temporal power sharing policies in cellular networks with energy constraints," *IEEE Trans. on Wireless Commun.*, vol. 11, no. 12, pp. 4405–4415, Dec. 2012.

[5] E. Oh, K. Son, and B. Krishnamachari, "Dynamic base station switching-on/off strategies for green cellular networks," *IEEE Trans. on Wireless Commun.*, vol. 12, no. 5, pp. 2126–2136, Mar. 2013.

[6] A. Abbasi and M. Ghaderi, "Energy cost reduction in celluar networks through dynamic base station activation," in *Proc. of SECON*, Singapore, Jul. 2014, pp. 363–371.

[7] K. Son, Y. Yi, and S. Chong, "Utility-optimal multi-pattern reuse in multi-cell networks," *IEEE Trans. on Wireless Commun.*, vol. 10, no. 1, pp. 142–153, Jan. 2011.

[8] E. Larsson, O. Edfors, F. Tufvesson, and T. Marzetta, "Massive MIMO for next generation wireless systems,," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 186–195, Feb. 2014.

[9] I. Hwang, B. Song, and S. Soliman, "A holistic view on hyper-dense heterogeneous and small cell networks," *IEEE Communications Magazine*, vol. 51, no. 6, pp. 20–27, Jun. 2013.

[10] J. Kim, H. Lee, and S. Chong, "TAES: Traffic-aware energy-saving base station sleeping and clustering in cooperative networks," in *Proc. of WiOpt*, Mumbai, India, May 2015, pp. 259–266.

[11] S. Han, C. I, Z. Xu, and C. Rowell, "Large scale antenna systems with hybrid analog and digital beamforming for millimeter wave 5g," *IEEE Communications Magazine*, vol. 53, no. 1, pp. 186–194, Jan. 2015.

[12] J. Kim, H. Lee, and S. Chong, "Virtual cell beamforming in cooperative networks," *IEEE JSAC*, vol. 32, no. 6, pp. 1126–1138, Jun. 2014.

[13] Y. Shi, J. Shang, and K. Letaief, "Group sparse beamforming for green Cloud-RAN," *IEEE Trans. on Wireless Commun.*, vol. 13, no. 5, pp. 2809–2823, Apr. 2014.

[14] M. Sharif and B. Hassibi, "On the capacity of MIMO broadcast channels with partial side information," *IEEE Trans. on Inform. Theory*, vol. 51, no. 2, pp. 506–522, Feb. 2005.

[15] M. Costa, "Writing on dirty paper (corresp.)," *IEEE Trans. on Inform. Theory*, vol. 29, no. 3, pp. 439–441, May 1983.

[16] M. Neely, "Stochastic network optimization with application to communication and queueing systems," *Synthesis Lectures on Communication Networks*, pp. 1–211, 2010.

[17] L. Kleinrock, *Queueing Systems*. Wiley, 1975.

[18] V. Singh, R. J. La, and M. A. Shayman, "Coordinated scheduling in mimo heterogeneous wireless networks using submodular optimization," in *Proc. of WiOpt*, May 2016, pp. 1–8.

[19] J. Kwak, Y. Kim, J. Lee, and S. Chong, "DREAM: Dynamic resource and task allocation for energy minimization in mobile cloud systems," *IEEE JSAC*, vol. 33, no. 12, pp. 2510–2523, Dec. 2015.

[20] A. Goldsmith, *Wireless communications*. Cambridge Univ. Press, 2005.

[21] M. Deruyck, W. Joseph, and L. Martens, "Power consumption model for macro and micro base stations," *Trans. on Emerging Telecommun. Techno.*, vol. 25, no. 3, pp. 320–333, Mar. 2014.

[22] J. Lee, V. S. Mirrokni, V. Nagarajan, and M. Sviridenko, "Non-monotone submodular maximization under matroid and knapsack constraints," in *Proc. of STOC*, New York, NY, USA, Jun. 2009, pp. 323–332.

[23] U. Feige, V. S. Mirrokni, and J. Vondrak, "Maximizing non-monotone submodular functions," *SIAM Journal on Computing*, vol. 40, no. 4, pp. 1133–1153, 2011.

[24] M. Kountouris, D. Gesbert, and T. Salzer, "Enhanced multiuser random beamforming: dealing with the not so large number of users case," *IEEE JSAC*, vol. 26, no. 8, pp. 1536–1545, Oct. 2008.

[25] Q. H. Spencer, A. L. Swindlehurst, and M. Haardt, "Zero-forcing methods for downlink spatial multiplexing in multiuser mimo channels," *IEEE Trans. on Signal Processing*, vol. 52, no. 2, pp. 461–471, 2004.

[26] G. Auer, O. Blume, V. Giannini, I. Godor, M. Imran, Y. Jading, E. Katranaras, M. Olsson, D. Sabella, P. Skillermark, I. Ali, and W. Wieslawa, "D2. 3: Energy efficiency analysis of the reference systems, areas of improvements and target breakdown," *EARTH*, pp. 1–60, 2010.