Energy Efficiency-Delay Tradeoff for a Cooperative NOMA System

Bing Ning, Wanming Hao, Aihua Zhang, Jiankang Zhang, Guan Gui

***Abstract*—The tradeoff between the energy efficiency (EE) and delay problem in cooperative relaying system is studied by using Non-orthogonal multiple access (NOMA) in this paper. To obtain an efficiency tradeoff between EE and delay, a stochastic-based EE optimization problem is formulated by considering the system queue stability. Then, the fractional programming and control parameter-based Lyapunov optimization method is proposed to solve the formulated problem. Furthermore, we derive the analytical bounds of EE and delay based on the control parameter. Finally, simulation results verify that the proposed cooperative NOMA system performs better than the traditional orthogonal multiple access (OMA) cooperative system.**

***Index Terms*—NOMA, cooperative relaying system, EE, delay**

I. Introduction

NOMA technique has recently been included into the 3GPP long term evolution advanced (LTE-A) standard, owing to its enormous potential in improving system spectrum efficiency [1]. Different from the traditional OMA, the users are allowed to share the same time/frequency resource in NOMA. Meanwhile, the success interference cancellation (SIC) is applied at the receivers to reduce the co-channel interference and extract desired components from the received signals.

Recently, cooperative NOMA is further proposed to improve the transmission reliability of the system, by exploiting the spatial diversity gain [2]-[6]. The power allocation strategy is optimized based on the closed form expressions for the base station's outage probability and sum rate in [2]. [3] derives the sum rate region in the cooperative NOMA system with compress-and-forward relaying by using the noisy network coding. In [4], the achievable rate is calculated approximatively through the Gauss-Chebyshev Integration method in a Rician fading channel. [5] proposes a two-stage selection strategy under different quality of services (QoSs) for the users, and then the closed-form expression of the outage probability is obtained. However, recent studies mainly focus on the rate and outage probability for the cooperative NOMA system.

The work was supported by the National Natural Science Foundation of China under Grant 61501530 and 61571401, Henan Educational Committee under Grant 16A510012,Henan science and technology planning project under Grant 182102210522 and Innovative Talent of Colleges and University of Henan Province under Grant 18HASTIT021(Corresponding author: W. Hao).

B. Ning and A. Zhang are with Zhongyuan University of Technology, School of electronic information, Zhengzhou, CN 450007.

W. Hao is with the School of Information Engineering, Zhengzhou University, Zhengzhou, Henan, 450001, China.

J. Zhang is with University of Southampton Faculty of Physical Sciences and Engineering, Southampton, UK SO17 1B.

G. Gui is with Nanjing University of Posts and Telecommunications Nanjing, CN 450001.

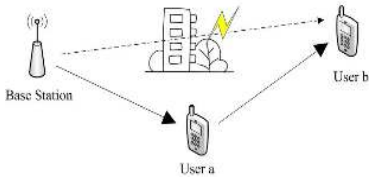

Fig. 1. The NOMA cooperative system model.

In fact, there lacks studies on delay, an important indicator for stabilization and reliability in cooperative systems. In addition, energy efficiency (EE), as a crucial criteria in the next generation wireless communication systems [7] [8], has not been investigated.

Towards filling the research gap, the EE-delay tradeoff problem for a cooperative NOMA system is studied, where a user with good channel condition acts as a relay to assist another user with bad channel condition. Specifically, we formulate a stochastic-based EE optimization problem by considering the system queue stability. Next, we transform the original formulated problem into two independent subproblems by using fractional programming and control parameter-based Lyapunove method, namely users (relay) and base satation transmission power optimization problem. The former belongs to a convex optimization problem, while the latter can also be transformed into a convex optimization problem by our proposed scheme. Next, we derive the analytical bounds of EE and delay, and then analyze the relation between them.

II. System Model and Problem formulation

A downlink cooperative NOMA system that consists of a base station (BS) and two pre-paired users, i.e., User a and User b is shown in Fig. 1. It is assumed that any direct link between the BS and User b doesn't exist due to the heavy shadowing or physical obstacles. As a result, User a has to act as relay for User b. To improve the transmission efficiency, User a operates on the full-duplex model, and the perfect self-interference cancellation is assumed to be available for User a. However, the focus of this letter is to investigate the fundamental tradeoff between EE and delay in a NOMA cooperative system. The developed analytical results in this letter not only provide insight for the case of perfect self-interference, but also can be used as a baseline for future research under imperfect self-interference. The slotted time mode is employed for the NOMA cooperative system with slots normalized to integral units, where slot τ refers to the time interval $[\tau, \tau+1), \tau \in \{0, 1, \cdots\}$. The data of User a or b randomly arrives at the BS in each slot, which are queued separately. As depicted in Fig. 2, $\mathbf{A}(\tau) = \{A_a(\tau), A_b(\tau)\}$ denotes

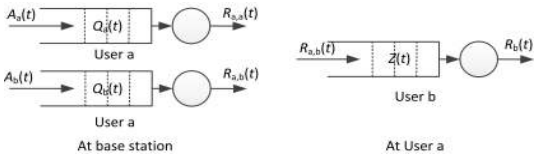

Fig. 2. The actual queuing process at BS and User a .

the process of random data arrivals of Users a and b, where $A_i(\tau)$ $(i \in \{a, b\})$ is independent and identically distributed (i.i.d.) over time with arrival rate λ_i, i.e., $E\{A_i(\tau)\} = \lambda_i$. $\mathbf{Q}(\tau) = \{Q_a(\tau), Q_b(\tau)\}$ denotes the number of data stored for Users a and b at BS at slot τ, whereas $Z(\tau)$ represents the number of data stored for User b at User a at slot τ.

The message of User a and b, $x_a(\tau)$ and $x_b(\tau)$, is superimposed at BS, as a downlink NOMA signal $x(\tau) = \sqrt{P_a(\tau)}x_a(\tau) + \sqrt{P_b(\tau)}x_b(\tau)$, where $P_a(\tau)$ and $P_b(\tau)$ denote the transmit power for $x_a(\tau)$ and $x_b(\tau)$, respectively. Then, the received signal of User a at slot τ can be formulated as

$$y_a(\tau) = g_a(\tau)x(\tau) + n_a(\tau), \tag{1}$$

where $g_a(\tau)$ represents the channel gain from the BS to User a, $n_a(\tau)$ follows $n_a(\tau) \sim CN(0, \delta^2)$. After receiving $y_a(\tau)$, User a first decodes $x_b(\tau)$ and then decodes its own signal $x_a(\tau)$ after removing $x_b(\tau)$. As a result, the SINR of $x_a(\tau)$ and $x_b(\tau)$ at User a are $\gamma_{a,a}(\tau) = \frac{P_a(\tau)|g_a(\tau)|^2}{\delta^2}$, $\gamma_{a,b}(\tau) = \frac{P_b(\tau)|g_a(\tau)|^2}{P_a(\tau)|g_a(\tau)|^2+\delta^2}$. Meanwhile, User a will transmit encoded $x_b(\tau)$ to User b, and the received signal at User b can be written as follows

$$y_b(\tau) = \sqrt{P_r(\tau)}g_b(\tau)x_b(\tau) + n_b(\tau), \tag{2}$$

where $P_r(\tau)$ denotes the transmit power of User a, $g_b(\tau)$ is the coefficient of channel fading from User a to User b, and $n_b(\tau)$ follows $n_b(\tau) \sim CN(0, \delta^2)$. The SINR of $x_b(\tau)$ at User b can be written as $\gamma_b(\tau) = P_r(\tau)|g_b(\tau)|^2/\delta^2$. Next, as depicted in Fig. 2, the queuing process at BS can be modeled as

$$Q_m(\tau + 1) = \max\left[Q_m(\tau) - R_{a,m}(\tau), 0\right] + A_m(\tau), m \in \{a, b\}, \tag{3}$$

where $R_{a,m}(\tau) = \log_2(1+\gamma_{a,m}(\tau))$, $R_{a,a}(\tau)$ and $R_{a,b}(\tau)$ denote the data rates of Users a and b at User a, respectively. Similarly, at User a, we have

$$Z(\tau + 1) = \max\left[Z(\tau) - R_b(\tau), 0\right] + R_{a,b}(\tau), \tag{4}$$

where $R_b(\tau) = \log_2(1 + \gamma_2(\tau))$ denotes the data rates of User b. Then, we model the total power consumption at BS and User a as $P_{\text{total}}(\tau) = \xi\left(P_a(\tau) + P_b(\tau) + P_r(\tau)\right) + P_C$, where ξ and P_C are the constants accounting for the inefficiency of the power amplify and the circuit power consumption at BS and User a, respectively. Accordingly, the sum rate can be represented as $R_{\text{sum}}(\tau) = R_{a,a}(\tau) + R_b(\tau)$. Next, we define the long-term EE [9] as

$$\eta_{\text{EE}} = \frac{\bar{R}_{\text{sum}}(\mathbf{P})}{\bar{P}_{\text{total}}(\mathbf{P})} = \frac{\lim_{T\to\infty}\frac{1}{T}\sum_{\tau=0}^{T-1}E[R_{\text{sum}}(\tau)]}{\lim_{T\to\infty}\frac{1}{T}\sum_{\tau=0}^{T-1}E[P_{\text{total}}(\tau)]}, \tag{5}$$

where $\mathbf{P} = \{P_a(\tau), P_b(\tau), P_r(\tau)\}\tau \in \{1, 2, \cdots\}$, and $E[\cdot]$ denotes the expectation.

Here, we focus on the EE problem in a steady-state network. The system queue stability means that all data can be transmitted to the users within the limited time. We define a single discrete-time queue $Q(\tau)$ $=$ $\limsup_{T\to\infty}\frac{1}{T}\sum_{\tau=0}^{T-1}E[Q(\tau)]$ $<$ ∞. Then, the system queue is strongly stable if all discrete-time queues are strongly stable. Since the average queue length is proportional to average delay, we can evaluate the system average delay by queue length, and then investigate the EE-delay tradeoff problem. Motivated by this, we formulate the following optimization problem

$$\max_{\{\mathbf{P}\}} \quad \eta_{\text{EE}} \tag{6a}$$

$$\text{s.t. } R_{a,m}(\tau) \geq R_m^{\min}, R_b(\tau) \geq R_b^{\min}, \forall m, \tag{6b}$$

$$\bar{P}_a + \bar{P}_b \leq P_{\text{BS}}^{\text{av}}, \bar{P}_r \leq P_r^{\text{av}}, \tag{6c}$$

$$P_a(\tau) + P_b(\tau) \leq P_{\text{BS}}^{\text{av}}, P_r(\tau) \leq P_r^{\text{av}}, \tag{6d}$$

$$\text{Queue } Q_m(\tau) \text{ and } Z(\tau) \text{ are mean rate stable}, \forall m, \tag{6e}$$

where $\bar{P}_m = \lim_{\tau\to\infty}\frac{1}{\tau}\sum_{t=0}^{\tau-1}E[P_m(t)]$ $(m \in \{a, b\})$ and $\bar{P}_r = \lim_{\tau\to\infty}\frac{1}{\tau}\sum_{t=0}^{\tau-1}E[P_r(t)]$. (6b) denotes the minimal rate requirements for Users a and b, and (6c) and (6d), respectively, denote the average and instantaneous power constraints for the BS and User a, whereas (6e) guarantees the stability of queues.

III. Problem Solution and Analysis

A. Problem Solution

It is obvious that the objective function is nonlinear fraction, thus (6) could not employ convex optimization to slove. According to fractional programming theory [10], the fractional objective function can be transformed into the subtract form, the optimal $\eta_{\text{EE}}^{\text{opt}}$ can be obtained if and only if

$$\max_{\{\mathbf{P}\in\mathbf{\Phi}\}}\{\bar{R}_{\text{sum}}(\mathbf{P}) - \eta_{\text{EE}}^{\text{opt}}\bar{P}_{\text{total}}(\mathbf{P})\} = \bar{R}_{\text{sum}}(\mathbf{P}^{\text{opt}}) - \eta_{\text{EE}}^{\text{opt}}\bar{P}_{\text{total}}(\mathbf{P}^{\text{opt}}) = 0, \tag{7}$$

where $\mathbf{\Phi}$ is the set of all feasible solutions. The detailed proof can be found in [10]. Therefore, the original problem (6) can be equivalently transformed into the following problem

$$\max_{\{\mathbf{P}\}} \bar{R}_{\text{sum}}(\mathbf{P}) - \eta_{\text{EE}}^{\text{opt}}\bar{P}_{\text{total}}(\mathbf{P}), \quad \text{s.t. (6b)} - \text{(6e)}. \tag{8}$$

It is still difficult to solve the above problem for these two reasons: 1) the time average expectations for the objective function, and 2) $\eta_{\text{EE}}^{\text{opt}}$ cannot be obtained in advance. Based on this, we define $\eta_{\text{EE}}(\tau)$ as

$$\eta_{\text{EE}}(\tau) = \frac{\sum_{t=0}^{\tau-1}R_{\text{sum}}(t)}{\sum_{t=0}^{\tau-1}P_{\text{total}}(t)}, \tau \in \{1, 2, \cdots\}, \tag{9}$$

where $\eta_{\text{EE}}(0) = 0$. Then, (6) can be transformed into the following by replacing $\eta_{\text{EE}}^{\text{opt}}$ with $\eta_{\text{EE}}(t)$ in (8) at each slot t

$$\max_{\{\mathbf{P}\}} \bar{R}_{\text{sum}}(\mathbf{P}) - \eta_{\text{EE}}(t)\bar{P}_{\text{total}}(\mathbf{P}), \quad \text{s.t. (6b)} - \text{(6e)}. \tag{10}$$

Due to the stochastic optimization problem in (10), we will propose an iterative method to solve it. According to the Lyapunov method [11], the average power constraints (6c) can be transformed into queue stability problems by defining virtual power queues as

$$O_a(\tau + 1) = \max\left[O_a(\tau) + P_a(\tau) + P_b(\tau) - P_{\text{BS}}^{\text{av}}, 0\right], \tag{11a}$$

$$O_b(\tau + 1) = \max\left[O_b(\tau) + P_r(\tau) - P_r^{\text{av}}, 0\right]. \tag{11b}$$

where $O_a(\tau)$ and $O_b(\tau)$, respectively, denote the virtual power queues at BS and User a. Next, let $\mathbf{\Omega}(\tau)$ =

$[Q_a(\tau), Q_b(\tau), Z(\tau), O_a(\tau), O_b(\tau)]$ denote the combined matrix of the traffic queues and virtual power queues. Accordingly, the Lyapunov function can be defined as

$$L(\mathbf{\Omega}(\tau)) = \sum\nolimits_{m\in\{a,b\}} (Q_m^2(\tau) + O_m^2(\tau))/2 + Z^2(\tau)/2. \quad (12)$$

Then, the upper bound of the Lyapunov drift-plus-penalty at slot τ can be writen as:

$$\begin{aligned} &E\{\Delta L(\mathbf{\Omega}(\tau))|\mathbf{\Omega}(\tau)\} + VE\{\eta_{\mathrm{EE}}(\tau)P_{\mathrm{total}}(\tau) - R_{\mathrm{sum}}(\tau)|\mathbf{\Omega}(\tau)\} \\ \le& B + E\{U(P_a(\tau), P_b(\tau), P_r(\tau))\} + Q_a(\tau)A_a(\tau) \\ &+Q_b(\tau)A_b(\tau) - O_a(\tau)P_{\mathrm{BS}}^{\mathrm{av}} - O_b(\tau)P_r^{\mathrm{av}}, \end{aligned} \quad (13)$$

where $U(P_a(\tau), P_b(\tau), P_r(\tau)) = V\eta_{\mathrm{EE}}(\tau)(P_a(\tau)+P_b(\tau)+P_r(\tau)) - V(R_{a,a}(\tau)+R_b(\tau)) - Q_a(\tau)R_{a,a}(\tau) - Q_b(\tau)R_{a,b}(\tau) + Z(\tau)(R_{a,b}(\tau) - R_b(\tau)) + O_a(\tau)(P_a(\tau)+P_b(\tau)) + O_b(\tau)P_r(\tau)$, $\Delta L(\mathbf{\Omega}(\tau)) = L(\mathbf{\Omega}(\tau+1)) - L(\mathbf{\Omega}(\tau))$. $V \ge 0$ is a correlation coefficient for controlling the tradeoff between the EE and delay, while B is a positive constant satisfying the following condition

$$B \ge E\{A_a^2(\tau) + A_b^2(\tau) + Z^2(\tau) + T_a^2(\tau) + T_b^2(\tau)|\mathbf{\Omega}(\tau)\}/2, \quad (14)$$

where $T_a(\tau) = P_a(\tau) + P_b(\tau) - P_{\mathrm{BS}}^{\mathrm{av}}$ and $T_b(\tau) = P_r(\tau) - P_r^{\mathrm{av}}$. Therefore, we can obtain the optimal power allocation in slot τ by minimizing the right-hand-side of (13) as follows

$$\min_{\{P_a(\tau),P_b(\tau),P_r(\tau)\}} \; U(P_a(\tau), P_b(\tau), P_r(\tau)) \;\; \text{s.t. (6b), (6d).} \quad (15)$$

(15) can be divided into two independent problems, i.e., the BS power optimization problem and User a power optimization problem. The BS power optimization problem can be formulated as

$$\begin{aligned} \min_{\{P_a(\tau),P_b(\tau)\}} \;& (V\eta_{\mathrm{EE}}(\tau) + O_a(\tau))(P_a(\tau)+P_b(\tau)) \\ &-(V + Q_a(\tau))R_{a,a}(\tau) - (Q_b(\tau) - Z(\tau))R_{a,b}(\tau) \quad (16a) \\ \text{s.t. }\; & R_{a,m}(\tau) \ge R_m^{\min}, P_a(\tau) + P_b(\tau) \le P_{\mathrm{BS}}^{\max}, \forall m. \quad (16b) \end{aligned}$$

Next, we propose a scheme to equivalently transform (16) in a convex optimization problem. According to the rate expression of $R_{a,a}(\tau)$ and $R_{a,b}(\tau)$, we have

$$P_a(\tau) = (2^{R_{a,a}(\tau)} - 1)\delta^2/|g_a(\tau)|^2, \quad (17a)$$

$$P_b(\tau) = (2^{R_{a,b}(\tau)} - 1)(P_a(\tau)|g_a(\tau)|^2 + \delta^2)/|g_a(\tau)|^2. \quad (17b)$$

In the expression of $P_b(\tau)$, we substitute $P_a(\tau)$ with $(2^{R_{a,a}(\tau)} - 1)\delta^2/|g_a(\tau)|^2$ and obtain $P_b(\tau) = 2^{R_{a,a}(\tau)}(2^{R_{a,b}(\tau)} - 1)\delta^2/|g_a(\tau)|^2$.

Finally, we transform (16) into the following problem

$$\begin{aligned} \min_{\{R_{a,a}(\tau),R_{a,b}(\tau)\}} \;& (V\eta_{\mathrm{EE}}(\tau) + O_a(\tau))f(R_{a,a}(\tau), R_{a,b}(\tau)) \\ &-(V + Q_a(\tau))R_{a,a}(\tau) - (Q_b(\tau) - Z(\tau))R_{a,b}(\tau) \quad (18a) \\ \text{s.t. }\; & R_{a,m}(\tau) \ge R_m^{\min}, \forall m, f(R_{a,a}(\tau), R_{a,b}(\tau)) \le P_{\mathrm{BS}}^{\max}, \quad (18b) \end{aligned}$$

where $f(R_{a,a}(\tau), R_{a,b}(\tau)) = P_a(\tau) + P_b(\tau) = (2^{(R_{a,a}(\tau)+R_{a,b}(\tau))} - 1)\delta^2/|g_a(\tau)|^2$. Since (18a) is convex w.r.t. $\{R_{a,a}(\tau), R_{a,b}(\tau)\}$, (18) is a convex optimization problem, which can be solved by standard convex technique (e.g., inter-point method). User a power optimization problem can be formulated as

$$\begin{aligned} \min_{\{P_r(\tau)\}} \;& V\eta_{\mathrm{EE}}(\tau)P_r(\tau) - (V + Z(\tau))R_b(\tau) + O_b(\tau)P_r(\tau) \quad (19a) \\ \text{s.t. }\; & R_b(\tau) \ge R_b^{\min}, P_r(\tau) \le P_r^{\max}. \quad (19b) \end{aligned}$$

Algorithm 1: Dynamic power allocation algorithm

Initialize $Q_m(0) \leftarrow 0$, $O_m(0) \leftarrow 0$, $Z(0) \leftarrow 0$, $m = \{a, b\}$ $\eta_{\mathrm{EE}}(0) \leftarrow 0$, $\tau \leftarrow 0$.
repeat
  Obtain the power $P_1(\tau)$ and $P_2(\tau)$ by solving (18).
  Obtain the power $P_r(\tau)$ by solving (19).
  Update $\tau \leftarrow \tau + 1$.
  Update queue length $Q_m(\tau)$ with (3).
  Update queue length $Z(\tau)$ with (4).
  Update virtual power queue length $O_m(\tau)$ with (11).
  Update $\eta_{\mathrm{EE}}(\tau)$ according to (9)
until $t = T$, where T is the total number of time slots;

It is obvious that (19) is a convex optimization problem, which can be solved by standard water-filling algorithm. Finally, we summarize the overall algorithm as Algorithm 1.

B. The Analysis of The EE and Delay

In the following, we analyze the relation between EE and delay. We assume that η_{EE}^* is the obtained optimal EE, and we have the following conditions

$$\begin{aligned} E\{R_{\mathrm{sum}}^{\phi}(\tau)\} &\ge E\{P_{\mathrm{total}}^{\phi}(\tau)(\eta_{\mathrm{EE}}^* - \omega)\}, \\ E\{P_a^{\phi}(\tau)+P_b^{\phi}(\tau)-P_{\mathrm{BS}}^{\mathrm{av}}|\mathbf{\Omega}(\tau)\} &= E\{P_a^{\phi}(\tau)+P_b^{\phi}(\tau)-P_{\mathrm{BS}}^{\mathrm{av}}\} \le \omega, \\ E\{P_r^{\phi}(\tau) - P_r^{\mathrm{av}}|\mathbf{\Omega}(\tau)\} &= E\{P_r^{\phi}(\tau) - P_r^{\mathrm{av}}\} \le \omega, \\ E\{R_{a,m}^{\phi}(\tau)-A_m(\tau)|\mathbf{\Omega}(\tau)\} &= E\{R_{a,m}^{\phi}(\tau)-A_m(\tau)\} \ge \pi, \\ E\{R_b^{\phi}(\tau)-R_{a,b}(\tau)|\mathbf{\Omega}(\tau)\} &= E\{R_b^{\phi}(\tau)-R_{a,b}(\tau)\} \ge \pi, \end{aligned} \quad (20)$$

where ω and π are positive constants, and ϕ denotes any feasible power allocation strategy. The detailed proof can be found in [11]. Then, we have

$$\begin{aligned} &E\{\Delta L(\mathbf{\Omega}(\tau))|\mathbf{\Omega}(\tau)\} + VE\{\eta_{\mathrm{EE}}(\tau)P_{\mathrm{total}}(\tau) - R_{\mathrm{sum}}(\tau)|\mathbf{\Omega}(\tau)\} \\ \le& B+Q_a(\tau)E\{A_a(\tau)-R_{a,a}^{\phi}(\tau)|\mathbf{\Omega}(\tau)\}+Q_b(\tau)E\{A_b(\tau)-R_{a,b}^{\phi}(\tau)|\mathbf{\Omega}(\tau)\} \\ &+Z(\tau)E\{R_{a,b}^{\phi}(\tau)-R_b^{\phi}(\tau)|\mathbf{\Omega}(\tau)\}+O_a(\tau)E\{P_a^{\phi}(\tau)+P_b^{\phi}(\tau)-P_{\mathrm{BS}}^{\mathrm{av}}|\mathbf{\Omega}(\tau)\} \\ &+O_b(\tau)E\{P_r^{\phi}(\tau)-P_r^{\mathrm{av}}|\mathbf{\Omega}(\tau)\}+VE\{\eta_{\mathrm{EE}}(\tau)P_{\mathrm{total}}^{\phi}(\tau)-R_{\mathrm{sum}}^{\phi}(\tau)|\mathbf{\Omega}(\tau)\} \end{aligned} \quad (21)$$

Plugging (20) into (21), taking $\omega \to 0$ and summing (21) with $\tau \in \{0, 1, \cdots, T-1\}$, we can obtain

$$\begin{aligned} &E\{L(\mathbf{\Omega}(T))\}-E\{L(\mathbf{\Omega}(0))\}+V\sum\nolimits_{\tau=0}^{T-1}(E\{\eta_{\mathrm{EE}}(\tau)P_{\mathrm{total}}(\tau)\}-E\{R_{\mathrm{sum}}(\tau)\}) \\ \le& TB+VE\{P_{\mathrm{total}}^{\phi}(\tau)\}\sum\nolimits_{\tau=0}^{T-1}E\{\eta_{\mathrm{EE}}(\tau)\} - VT\eta_{\mathrm{EE}}^*E\{P_{\mathrm{total}}^{\phi}(\tau)\} \\ &-\pi\sum\nolimits_{\tau=0}^{T-1}(Q_a(\tau)+Q_b(\tau)+Z(\tau)). \end{aligned} \quad (22)$$

Dividing (22) by $T\pi$ and taking a limit as $T \to \infty$, rearranging terms with the fact that $E\{L(\mathbf{\Omega}(T))\} < \infty$ yields

$$\begin{aligned} &\lim_{T\to\infty}\frac{1}{T}\sum\nolimits_{\tau=0}^{T-1}(Q_a(\tau)+Q_b(\tau)+Z(\tau)) \\ \le& \frac{B}{\pi}+\frac{V}{\pi}E\{P_{\mathrm{total}}^{\phi}(\tau)\}\lim_{T\to\infty}\frac{1}{T}\sum_{\tau=0}^{T-1}E\{\eta_{\mathrm{EE}}(\tau)\}+\frac{V}{\pi}\lim_{T\to\infty}\frac{1}{T}\sum_{\tau=0}^{T-1}E\{R_{\mathrm{sum}}(\tau)\} \\ \le& \frac{B + VP_{\mathrm{total}}^{\max}\eta_{\mathrm{EE}}^{\max} + VR_{\mathrm{sum}}^{\max}}{\pi}, \end{aligned} \quad (23)$$

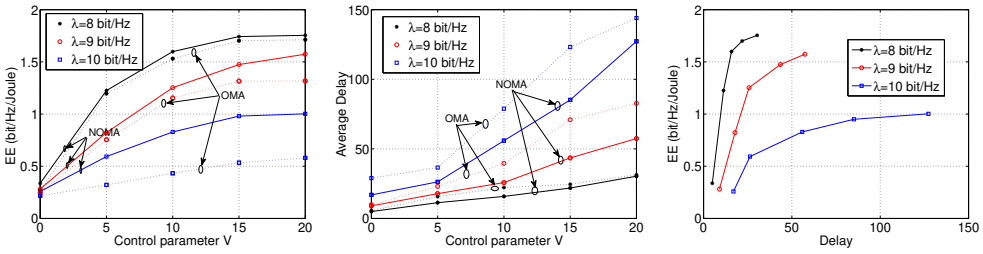

Fig. 3. (a) EE versus V. (b) Average delay versus V. (c) EE versus Delay.

where $\lim_{T\to\infty}\frac{1}{T}\sum_{\tau=0}^{T-1}(Q_a(\tau)+Q_b(\tau)+Z(\tau))$ denotes the average queue length, while the last step follows $E\{P^{\phi}_{\text{total}}(\tau)\}\le P^{\max}_{\text{total}}$, $E\{R_{\text{sum}}(\tau)\}\le R^{\max}_{\text{sum}}$, and $E\{\eta_{\text{EE}}(\tau)\}\le\eta^{\max}_{\text{EE}}$ with $P^{\max}_{\text{total}}$, $R^{\max}_{\text{sum}}$ and $\eta^{\max}_{\text{EE}}$ being some finite constants.

Since the last term of (22) is positive, dividing (22) by VT and taking a limit as $T\to\infty$, we can obtain

$$\frac{B}{V}+E\{P^{\phi}_{\text{total}}(\tau)\}\lim_{T\to\infty}\frac{1}{T}\sum\nolimits_{\tau=0}^{T-1}E\{\eta_{\text{EE}}(\tau)\}-\eta^{*}_{\text{EE}}E\{P^{\phi}_{\text{total}}(\tau)\}\ge 0, \tag{24}$$

where the zero at the right-hand side is due to $\lim_{T\to\infty}\frac{1}{T}\sum_{\tau=0}^{T=1}(E\{\eta_{\text{EE}}(\tau)P_{\text{total}}(\tau)\}-E\{R_{\text{sum}}(\tau)\})=0$. We have

$$\eta_{\text{EE}}=\lim_{T\to\infty}\frac{1}{T}\sum\nolimits_{\tau=0}^{T-1}E\{\eta_{\text{EE}}(\tau)\}\ge\eta^{*}_{\text{EE}}-\frac{B}{VE\{P^{\phi}_{\text{total}}(\tau)\}}\ge\eta^{*}_{\text{EE}}-\frac{B}{VP^{\min}_{\text{total}}} \tag{25}$$

We can observe that EE increases with V. Since the average queue length also grows with V (as shown in (23)), increase in V also brings larger delay. Therefore, there is a tradeoff between EE and delay, and it is important to choose a proper V to obtain the required performance in realistic systems.

IV. Simulation Results

Simulation results are provided to show the EE and delay of the proposed algorithm. For comparison, we also provide the results for an OMA system, where the time division multiple access (TDMA) is adopted. For the TDMA cooperative system, two time slots are needed. In the first time slot, the BS only serves User a while in the second time slot, User a acts as a full-duplex relay for the transmission from the BS to User b. Then, we formulate the same optimization problem and apply the same algorithm as in NOMA system to obtain the EE and delay of the TDMA cooperative system. The path-loss model between two nodes is given by $d^{-\alpha}$, where d denotes the distance between the transmitter and the receiver and α is the path-loss exponent. The fast-fading coefficients are all generated as i.i.d. Rayleigh random variables with unit variances. We assume that the distance between BS and User a is 100 m, whereas that between User a and User b is 20 m. We set the minimal rate requirement of Users a and b to 2 bit/Hz. The noise power is -100 dBm, and the pass loss exponent is 3.8. $P_C=6$ W, $P^{\text{av}}_{\text{BS}}=P^{\max}_{\text{BS}}=46$ dBm, $P^{\text{av}}_{r}=P^{\max}_{r}=20$ dBm, and $\xi=0.38$. The arrival rate for Users a and b in each time slot t is assumed to be uniformly distributed in $[0,2\lambda]$, i.e., $\lambda_a=\lambda_b=\lambda$. 10,000 slots is used to approximate $t\to\infty$.

Fig. 3(a) shows the EE versus V under different arrival rates. It is clear that the EE first increases with V and then tends to become stable. In addition, the EE is higher under lower arrival rates. This is because that the BS and User a need to consume more power to compensate for the delay under higher arrival rates. Moreover, we can see that NOMA outperforms OMA in term of the EE. Fig. 3(b) shows the average delay versus V under different arrival rates. It can be observed that the average delay increases with V. Meanwhile, we find that the average delay increases with the arrival rates which is due to that a higher arrival rate contributes to a larger delay for a given transmit power. Also, the average delay is lower under NOMA than that under OMA. Fig. 3(c) shows the relation between EE and delay under NOMA scheme. It is clear that the EE increases with delay, which indicates that a high EE can only be achieved at the cost of large delay, and vice verse.

V. Conclusions

In this letter, we investigated the EE and delay tradeoff in a NOMA cooperative system. We formed a long-term EE maximization problem under guaranteeing the stability of the system. The formulated problem was solved by the Lyapunov optimization approach. On this basis, we analyzed the EE-delay tradeoff. The results demonstrated that the performance of the NOMA cooperative system is better than OMA one.

References

[1] S. M. R. Islam et al., "Resource Allocation for Downlink NOMA Systems: Key Techniques and Open Issues," *IEEE Wireless Commu.*, vol. 25, no. 2, pp. 40-47, April 2018.

[2] L. Zhang et al., "Performance analysis and optimization in downlink NOMA systems with cooperative full-duplex relaying," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2398-2412, Oct. 2017.

[3] J. So et al., "Improving non-orthogonal multiple access by forming relaying broadcast channels," *IEEE Commun. Lett.*, vol. 20, no. 9, pp. 1816-1819, Sep. 2016.

[4] R. Jiao et al., "On the performance of NOMA-based cooperative relaying systems over rician fading channels," *IEEE Trans. Veh. Tech.*, to appear.

[5] Z. Yang et al., "Novel relay selection strategies for cooperative NOMA," *IEEE Trans. Veh. Tech.*, vol. 66, no. 11, pp. 10114-10123, Nov. 2017.

[6] M. Zeng et al., "Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster," *IEEE Journal on Selected Areas in Commun.*, vol. 35, no. 10, pp. 2413-2424, July 2017.

[7] Z. Chang et al, "Collaborative Mobile Clouds: An Energy Efficient Paradigm for Content Sharing," *IEEE Wireless Commu.*, to appear.

[8] W. Yu et al., "Tradeoff analysis and joint optimization of link-layer energy efficiency and effective capacity toward green communications," *IEEE Wireless Commu.*, vol. 15, no. 5, pp. 3339-3353, May 2016.

[9] Y. Li et al., "Green Heterogeneous Cloud Radio Access Networks: Potential Techniques, Performance Tradeoffs and Challenges," *IEEE Commun. Mag.*, vol. 55, no. 11, pp. 33-39, Nov. 2017.

[10] A., Zappone et al., "Energy efficiency in wireless networks via fractional programming theory," Now publisher, Foundations and Trends in Communications and Information Theory, pp. 185-396, 2015.

[11] M. J. Neely, "Dynamic optimization and learning for renewal systems," *IEEE Trans. Autom. Control,* vol. 58, no. 1, pp. 3246, Jan. 2013.