Cache-Aided Non-Orthogonal Multiple Access

Lin Xiang, Derrick Wing Kwan Ng, Xiaohu Ge,
Zhiguo Ding, Vincent W.S. Wong, and Robert Schober

***Abstract*—In this paper, we propose a novel joint caching and non-orthogonal multiple access (NOMA) scheme to facilitate advanced downlink transmission for next generation cellular networks. In addition to reaping the conventional advantages of caching and NOMA transmission, the proposed cache-aided NOMA scheme also exploits cached data for interference cancellation which is not possible with separate caching and NOMA transmission designs. Furthermore, as caching can help to reduce the residual interference power, several decoding orders are feasible at the receivers, and these decoding orders can be flexibly selected for performance optimization. We characterize the achievable rate region of cache-aided NOMA and investigate its benefits for minimizing the time required to complete video file delivery. Our simulation results reveal that, compared to several baseline schemes, the proposed cache-aided NOMA scheme significantly expands the achievable rate region for downlink transmission, which translates into substantially reduced file delivery times.**

I. Introduction

Wireless caching is a content-centric networking solution to meet the large downlink capacity demands introduced by video streaming in fifth generation (5G) cellular networks [1]–[6]. Recently, caching at streaming user equipments (UEs), e.g. smartphones and tablets, has been advocated [7], [8] to enhance the streaming quality of experience (QoE) while reducing (i.e., *offloading*) over-the-air traffic. This poses significant challenges for the design of cache placement and delivery as the aggregate cache capacity is distributed across non-cooperating devices with small individual cache memory sizes. Besides, the actual requests of the UEs are difficult to predict during cache placement due to the users' mobility and the random nature of the users' requests.

In the literature, *coded caching* has been proposed as an effective solution for caching at UEs [7], [8]. By exploiting the cached data as side information, a *coded multicast* format is created for simultaneous error-free video delivery to multiple users, which leads to a multiplicative performance gain that scales with the aggregate cache memory size of the UEs. However, coded caching has an exponential-time computational complexity. Moreover, the caching concepts proposed in [7], [8] are mainly applicable to noiseless and error-free communication links, e.g. in wireline networks. For wireless networks impaired by fading and noise, however, the performance of coded multicast is limited by the weakest user with the poorest channel condition within the multicast group.

On the other hand, non-orthogonal multiple access (NOMA) is an efficient approach for wireless multiuser transmission that alleviates the adverse effects of fading [9], [10]. Different from multicast and coded multicast, NOMA pairs multiple simultaneous downlink transmissions on the same time-frequency resource via power domain or code domain multiplexing [11]. Strong users with favorable channel conditions can cancel the interference caused by weak users with poor channel conditions that are paired on the same time-frequency resource, and hence, achieve a high data rate at low transmit powers. Therefore, high transmit powers can be allocated to weak users to achieve communication fairness [12]. NOMA has also been extended to multicarrier and multi-antenna systems; see [13], [14] and references therein.

So far, wireless caching and NOMA were either investigated separately or combined in a straightforward manner [15]. For the latter case, NOMA was shown to improve the performance of both caching and delivery in [15]. In this paper, however, the joint design of caching and NOMA is advocated to maximize the performance gains introduced by caching at UEs. We show that the joint design of caching and NOMA can significantly outperform the straightforward combination of caching and NOMA. To this end, we consider a simple distributed caching scheme for video file delivery. By splitting the video files into several subfiles, superposition transmission of the requested uncached subfiles is enabled during delivery. If the cached content is *hit*, i.e., requested by the caching UE, the proposed cache-aided NOMA scheme enables traditional offloading of the video traffic. Otherwise, the *missed* cached data, which is not requested by the caching UE, is still exploitable as side information to facilitate (partial) interference cancellation for NOMA. The resulting cache-enabled interference cancellation (CIC) can neither be exploited by separate caching and NOMA designs nor by the scheme in [15].

With CIC, cached data is useful during file delivery even if the users' requests cannot be accurately predicted *a priori*. Moreover, joint CIC and successive interference cancellation (SIC) improves the interference mitigation capability at the UEs and increases the number of possible decoding orders compared to conventional NOMA. Furthermore, the performance of both strong and weak users can benefit from CIC. However, adaptive adjustment of the decoding order according to the cache and channel statuses is critical for reaping the benefits of cache-aided NOMA. Hence, we investigate the joint decoding order selection and power and rate allocation optimization problem for minimization of the file delivery time for *fast* video delivery. The main contributions of this paper are as follows:

- We propose a novel cache-aided NOMA delivery scheme for spectrally efficient downlink transmission. Thereby, cached data is exploited for cancellation of NOMA interference. We characterize the achievable rate region of the proposed scheme.
- We jointly optimize the NOMA decoding order and the rate and power allocations for minimization of the delivery time. As the formulated optimization problem is nonconvex, we propose an iterative method to solve it optimally via solving a sequence of convex problems.
- We show by simulation that the proposed scheme leads to a considerably larger achievable rate region and a significantly reduced delivery time compared to several baseline schemes, including the straightforward combination of caching and NOMA.

Notations: $\mathbb{C}$ and $\mathbb{R}_+$ denote the sets of complex and nonnegative real numbers, respectively. $\mathbb{E}(\cdot)$ is the expectation operator. $\mathcal{CN}\left(\mu, \sigma^2\right)$ represents the complex Gaussian distri-

bution with mean μ and variance σ^2. $\mathbf{1}[\cdot]$ denotes an indicator function which is 1 when the event is true and 0 otherwise. For decoding the received signals, the notation $i \overset{(n)}{\rightarrow} x_f$ means that x_f is the nth decoded signal at UE i. Similarly, $i \overset{(n)}{\rightarrow} (x_f, x_{f'})$ means that signals x_f and $x_{f'}$ are jointly decoded in the nth decoding step. Finally, $C(\Gamma) \triangleq \log_2(1+\Gamma)$ denotes the capacity function of an additive white Gaussian noise (AWGN) channel, where Γ is the signal-to-interference-plus-noise ratio (SINR).

II. System Model

We consider cellular video streaming from a base station (BS) to two UEs indexed by i and j, respectively. The BS and the UEs have a single antenna, respectively. UEs i and j request files W_A and W_B of sizes V_A and V_B bits, respectively, where $W_A \neq W_B$. The respective requests are denoted as (i, A) and (j, B). Each UE is equipped with a cache of size C_k bits. Thereby, UE $k \in \{i, j\}$ can place portions of file $f \in \{A, B\}$ into its cache prior to the time of request, e.g. during the early mornings when cellular traffic is low. As the cache placement is completed before the users' requests are known, the users may cache files which they later do not request. We assume that UE k, $k \in \{i, j\}$, has cached $c_{kf} \in [0, 1]$ portion of file W_f, $f \in \{A, B\}$.

A. Cache Status and File Splitting

Let us define the minimum and maximum portions of content cached for file W_f by

$$\underline{c}_f \triangleq \min_{k\in\{i,j\}} c_{kf} \quad \text{and} \quad \overline{c}_f \triangleq \max_{k\in\{i,j\}} c_{kf}, \tag{1}$$

which correspond to the cache status at user

$$\underline{k}_f \triangleq \arg\min_{k\in\{i,j\}} c_{kf} \ \text{ and } \ \overline{k}_f \triangleq \arg\max_{k\in\{i,j\}} c_{kf}, \tag{2}$$

$f \in \{A, B\}$, respectively. Based on (1) and (2), four cache configurations are possible at the time of request:

Case I: $i = \overline{k}_B$ and $j = \overline{k}_A$, i.e., $i = \underline{k}_A$ and $j = \underline{k}_B$;
Case II: $i = \overline{k}_B$ and $j = \underline{k}_A$, i.e., $i = \overline{k}_A$ and $j = \underline{k}_B$;
Case III: $i = \underline{k}_B$ and $j = \overline{k}_A$, i.e., $i = \underline{k}_A$ and $j = \overline{k}_B$;
Case IV: $i = \underline{k}_B$ and $j = \underline{k}_A$, i.e., $i = \overline{k}_A$ and $j = \overline{k}_B$.

In particular, Case I reflects the scenario where the non-requesting user has cached a larger portion of file W_f than the requesting user, which constitutes an unfavorable cache placement for both users but cannot be avoided in practice as user requests cannot be predicted accurately. In the following, due to the limited space, we only consider Case I. However, the derivations for Case I can be extended to Cases II–IV in a relatively straightforward manner.

Let $Z_k \triangleq (Z_{k,A}, Z_{k,B})$, $k \in \{i, j\}$, denote the cache status of UE k, where $Z_{k,f}$, $f \in \{A, B\}$, is the cached content of file W_f. We assume that the video data of file W_f is sequentially organized. Moreover, based on different user requests and cache configurations, W_f is split into three subfiles (W_{f0}, W_{f1}, W_{f2}) for adaptive file delivery. As illustrated in Fig. 1, W_{f0} and W_{f2} of size $\underline{c}_f V_f$ and $(1-\overline{c}_f)V_f$ bits are the video chunks which are cached and uncached at both UEs, respectively, whereas subfile W_{f1} of size $(\overline{c}_f - c_{kf})V_f$ bits is only cached at UE $\overline{k}_f$. Hence, we have $Z_{\underline{k}_f,f} = (W_{f0})$ and $Z_{\overline{k}_f,f} = (W_{f0}, W_{f1})$, $f \in \{A, B\}$. As the cached data $Z_{\underline{k}_f,f}$ is the prefix of $Z_{\overline{k}_f,f}$, the considered caching scheme is referred to as *prefix caching* in [16].

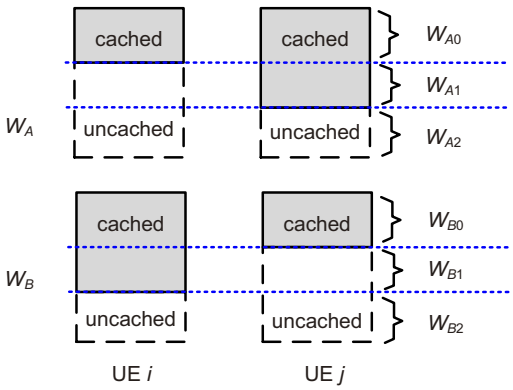

Fig. 1. Illustration of file splitting for cache-aided NOMA, assuming the cache configuration in Case I.

B. NOMA Transmission

For video delivery, we assume a frequency flat quasi-static fading channel, where the channel coherence time exceeds the time needed for completion of file delivery. The received signal at UE k is given by

$$y_k = h_k x + z_k, \tag{3}$$

where $h_k \in \mathbb{C}$ denotes the channel gain between the BS and UE k, and is constant during the transmission of file W_f. x is the transmit signal and $z_k \sim \mathcal{CN}(0, \sigma_k^2)$ is the AWGN at UE k.

The BS is assumed to know the cache statuses Z_i and Z_j during video delivery. Hence, the BS only transmits the uncached subfiles requested by the UEs. Thereby, for file W_f, subfiles W_{f1}, W_{f2}, $f \in \{A, B\}$, are encoded employing four independent codebooks at the BS and the corresponding codewords are superposed before being broadcasted over the channel according to the NOMA principle. The resulting BS transmit signal is given by

$$x = \sqrt{p_{i,1}}x_{A1} + \sqrt{p_{i,2}}x_{A2} + \sqrt{p_{j,1}}x_{B1} + \sqrt{p_{j,2}}x_{B2}, \tag{4}$$

where x_{fs}, $f \in \{A, B\}$, $s \in \{1, 2\}$, is the codeword corresponding to subfile W_{fs}, and $\mathbb{E}\big[|x_{fs}|^2\big] = 1$. Furthermore, $p_{k,s} \geq 0$, $k \in \{i, j\}$, $s \in \{1, 2\}$, denotes the transmit power of x_{fs}.

As the channel is static, we consider time-invariant power allocation, i.e., the powers, $p_{k,s}$, are fixed during file delivery. The total transmit power at the BS is constrained to P, i.e.,

$$\text{C1: } \sum\nolimits_{k\in\{i,j\}} \sum\nolimits_{s\in\{1,2\}} p_{k,s} \leq P. \tag{5}$$

We define $\mathbf{p} \triangleq (p_{i,1}, p_{i,2}, p_{j,1}, p_{j,2})$ and $\mathcal{P} \triangleq \{\mathbf{p} \in \mathbb{R}_+^4 \mid \text{C1}\}$ as the power allocation vector and the corresponding feasible set, respectively.

C. Joint CIC and SIC Decoding

The proposed cache-aided NOMA scheme enables CIC at the receiver, which is not possible for conventional NOMA. The joint CIC and SIC receiver performs CIC preprocessing of the received signal before SIC decoding as illustrated in Fig. 2. In particular, the interference caused by transmit signal x_{A1} (x_{B1}), which is requested by UE i (UE j), can be canceled at UE j (UE i) by exploiting the cached data $Z_{j,A}$ ($Z_{i,B}$). Hence, the residual received signal after CIC

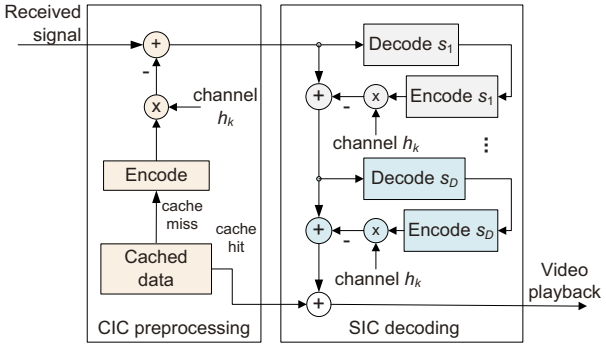

Fig. 2. Joint CIC and SIC decoding at receiver $k \in \{i, j\}$ for cache-aided NOMA. $s_1, \ldots, s_D$, $D \leq 4$, represent the residual signals x_{fs}, $f \in \{A, B\}$, $s \in \{1, 2\}$, which are not canceled by CIC but decoded successively by employing SIC.

preprocessing is given by

$$y_i^{\text{CIC}} = h_i(\sqrt{p_{i,1}}x_{A1} + \sqrt{p_{i,2}}x_{A2} + \sqrt{p_{j,2}}x_{B2}) + z_i, \quad (6)$$

$$y_j^{\text{CIC}} = h_j(\sqrt{p_{i,2}}x_{A2} + \sqrt{p_{j,1}}x_{B1} + \sqrt{p_{j,2}}x_{B2}) + z_j. \quad (7)$$

Remark 1. For the proposed cache-aided NOMA scheme, c_{kf} portion of file W_f is not transmitted at all, while $\bar{c}_f - c_{kf}$ portion of file W_f can be removed from the received signal of the non-requesting UE k' via CIC, where $(k, f) \in \{(i, A), (j, B)\}$ and $k' \neq k$. As such, the proposed scheme can exploit $\bar{c}_f$ portion of W_f for performance improvement, even for the unfavorable cache configuration of Case I, whereas a straightforward combination of caching and NOMA can only exploit the cached portion c_{kf} of the requested file [15].

As CIC reduces the multiuser interference power, multiple decoding orders become possible for SIC processing of y_k^{CIC}. For example, there are $4! = 24$ possible decoding orders based on (6) and (7), compared to $2! = 2$ for conventional NOMA. This leads to a substantially increased flexibility in decoding the video data based on y_k^{CIC}.

Optimizing the SIC decoding order and the respective transmission rates and power allocations based on the cache status and channel conditions enhances the performance of video delivery. On the other hand, decoding order optimization is a combinatorial problem, which may increase complexity. However, by careful inspection of the SIC decoding conditions, we show in Section III that the optimal decoding order is contained in a small subset of all possible decoding orders, and hence the associated complexity is limited.

III. Achievable Rate Region and Delivery Time Minimization

In this section, we evaluate the achievable rate region of the proposed cache-aided NOMA scheme. Based on the derived results, we then minimize the delivery time during file transfer by optimizing the decoding order and the power and rate allocation. Let $\mathbf{r} \triangleq (r_{i,1}, r_{i,2}, r_{j,1}, r_{j,2})$ be the rate allocation vector, where $r_{k,s} \geq 0$ is the rate for delivering x_{fs} to UE k, $(k, f) \in \{(i, A), (j, B)\}$, $s \in \{1, 2\}$. We define $\alpha_k \triangleq \frac{\sigma_k^2}{|h_k|^2}$, $k \in \{i, j\}$, as the effective noise variance at UE k. Without loss of generality, we assume $\alpha_i < \alpha_j$, i.e., UE i has a larger channel gain than UE j.

A. Derivation of Achievable Rate Region

According to (6) and (7), two subfiles, x_{f1} and x_{f2}, are delivered to each user and x_{B1} (x_{A1}) is canceled at UE i (j) by CIC. Moreover, x_{A2} and x_{B2}, which are interference signals at one user, are commonly received at both users, whereas x_{A1} and x_{B1} are received only at the requesting users. The interference signals can be decoded and canceled only if the SIC decoding condition is fulfilled, i.e., the received SINR for x_{A2} and x_{B2} at the non-requesting users, UE j and UE i, has to exceed that at the requesting users, UE i and UE j, respectively. In contrast, signals x_{A1} and x_{B1} can be decoded without such constraint. Depending on which signal is decoded first, three cases can be distinguished: for the first two cases, signals x_{A1} and x_{B1} are decoded first at the requesting users, respectively, whereas, for the third case, the interference signals x_{A2} and x_{B2} are decoded first. For these cases, as (6) and (7) constitute a non-degraded broadcast channel, the corresponding achievable rate regions have to be evaluated for specific power regions individually.

1) UE $i \overset{(1)}{\rightarrow} x_{A1}$: If UE i decodes x_{A1} first, signals $y_i^{(1)} = h_i(\sqrt{p_{i,2}}x_{A2} + \sqrt{p_{j,2}}x_{B2}) + z_i$ and $y_j^{(1)} = h_j(\sqrt{p_{i,2}}x_{A2} + \sqrt{p_{j,2}}x_{B2} + \sqrt{p_{j,1}}x_{B1}) + z_j$ have to be decoded subsequently. The achievable rate region is provided in Proposition 1.

Proposition 1. When UE $i \overset{(1)}{\rightarrow} x_{A1}$, the rate region $\mathcal{R}_1(\mathcal{P}_1) \bigcup \mathcal{R}_2(\mathcal{P}_2)$ is achievable, where

$$\mathcal{R}_1(\mathcal{P}_1) \triangleq \bigcup_{\mathbf{p} \in \mathcal{P}_1} \left\{ \mathbf{r} \;\middle|\; \begin{array}{l} r_{i,1} \leq C_{i,1} = C\left(\frac{p_{i,1}}{p_{i,2}+p_{j,2}+\alpha_i}\right) \\ r_{i,2} \leq C_{i,2} = C\left(\frac{p_{i,2}}{\alpha_i}\right) \\ r_{j,s} \leq C_{j,s} = C\left(\frac{p_{j,s}}{p_{i,2}+\alpha_j}\right), s = 1, 2 \\ r_{j,1} + r_{j,2} \leq C_{j,1,2} = C\left(\frac{p_{j,1}+p_{j,2}}{p_{i,2}+\alpha_j}\right) \end{array} \right\}$$

$$\mathcal{R}_2(\mathcal{P}_2) \triangleq \bigcup_{\mathbf{p} \in \mathcal{P}_2} \left\{ \mathbf{r} \;\middle|\; \begin{array}{l} r_{i,1} \leq C_{i,1} = C\left(\frac{p_{i,1}}{p_{i,2}+p_{j,2}+\alpha_i}\right) \\ r_{i,2} \leq C_{i,2} = C\left(\frac{p_{i,2}}{p_{j,2}+\alpha_i}\right) \\ r_{j,1} \leq C_{j,1} = C\left(\frac{p_{j,1}}{\alpha_j}\right) \\ r_{j,2} \leq C_{j,2} = C\left(\frac{p_{j,2}}{p_{i,2}+p_{j,1}+\alpha_j}\right) \end{array} \right\}$$

with $\mathcal{P}_1 = \mathcal{P}$ and $\mathcal{P}_2 \triangleq \{\mathbf{p} \in \mathcal{P} \mid p_{j,2} - p_{j,1} > \alpha_j - \alpha_i\}$. For $\mathcal{R}_1(\mathcal{P}_1)$, the decoding orders for UEs i and j are given as $i \overset{(1)}{\rightarrow} x_{A1} \overset{(2)}{\rightarrow} x_{B2} \overset{(3)}{\rightarrow} x_{A2}$ and $j \overset{(1)}{\rightarrow} (x_{B1}, x_{B2})$, respectively. For $\mathcal{R}_2(\mathcal{P}_2)$, the decoding orders are $i \overset{(1)}{\rightarrow} x_{A1} \overset{(2)}{\rightarrow} x_{A2}$ and $j \overset{(1)}{\rightarrow} x_{A2} \overset{(2)}{\rightarrow} x_{B2} \overset{(3)}{\rightarrow} x_{B1}$, respectively.

Proof: Please refer to Appendix A. ∎

Remark 2. In Proposition 1, the interference for decoding x_{A2} is reduced after x_{A1} has been decoded and canceled from y_i^{CIC}. Hence, decoding x_{A1} first is desirable when e.g. W_{A1} has a smaller size and/or requires a lower delivery rate than W_{A2}. The decoding orders for $\mathcal{R}_1(\mathcal{P}_1)$ favor the delivery of x_{A2} to UE i as it experiences no interference after SIC, and thus can attain a high data rate $r_{i,2}$ even for small transmit powers $p_{i,2}$. In contrast, the decoding orders for $\mathcal{R}_2(\mathcal{P}_2)$ favor the delivery of x_{B1} to UE j. This implies that $\mathcal{R}_2(\mathcal{P}_2)$ expands $\mathcal{R}_1(\mathcal{P}_1)$ along $r_{j,1}$.

2) UE $j \overset{(1)}{\rightarrow} x_{B1}$ excluding UE $i \overset{(1)}{\rightarrow} x_{A1}$: [1]Decoding and canceling x_{B1} first improves the SINR of x_{B2} at UE j, which is desirable when subfile W_{B1} has a smaller size than W_{B2}. The resulting signals after x_{B1} has been canceled are $y_i^{(1)} = h_i(\sqrt{p_{i,1}}x_{A1} + \sqrt{p_{i,2}}x_{A2} + \sqrt{p_{j,2}}x_{B2}) + z_i$ and $y_j^{(1)} = h_j(\sqrt{p_{i,2}}x_{A2} + \sqrt{p_{j,2}}x_{B2}) + z_j$. The corresponding achievable rate region is given in Proposition 2.

[1]The achievable rate region for UE $j \overset{(1)}{\rightarrow} x_{B1}$ and UE $i \overset{(1)}{\rightarrow} x_{A1}$ is already included in $\mathcal{R}_1(\mathcal{P}_1)$, and hence, excluded herein.

Proposition 2. When UE $j \overset{(1)}{\rightarrow} x_{B1}$ but UE $i \overset{(1)}{\rightarrow} x_{A1}$ is excluded, the achievable rate region is given by $\mathcal{R}_3(\mathcal{P}_3) \bigcup \mathcal{R}_4(\mathcal{P}_4)$, where

$$\mathcal{R}_3(\mathcal{P}_3) \triangleq \bigcup_{\mathbf{p}\in\mathcal{P}_3} \left\{ \mathbf{r} \left| \begin{array}{l} r_{i,s} \le C\left(\frac{p_{i,s}}{\alpha_i}\right),\ s=1,2 \\ r_{i,1}+r_{i,2} \le C\left(\frac{p_{i,1}+p_{i,2}}{\alpha_i}\right) \\ r_{j,1} \le C\left(\frac{p_{j,1}}{p_{i,2}+p_{j,2}+\alpha_j}\right) \\ r_{j,2} \le C\left(\frac{p_{j,2}}{p_{i,2}+\alpha_j}\right) \end{array} \right. \right\}$$

$$\mathcal{R}_4(P_4) \triangleq \bigcup_{\mathbf{p}\in\mathcal{P}_4} \left\{ \mathbf{r} \left| \begin{array}{l} r_{i,1} \le C\left(\frac{p_{i,1}}{p_{j,2}+\alpha_i}\right) \\ r_{i,2} \le C\left(\frac{p_{i,2}}{p_{i,1}+p_{j,2}+\alpha_i}\right) \\ r_{j,1} \le C\left(\frac{p_{j,1}}{p_{i,2}+p_{j,2}+\alpha_j}\right) \\ r_{j,2} \le C\left(\frac{p_{j,2}}{\alpha_j}\right) \end{array} \right. \right\}$$

with $\mathcal{P}_3 \triangleq \{\mathbf{p}\in\mathcal{P} \mid p_{i,1} < \alpha_j - \alpha_i\}$ and $\mathcal{P}_4 = \mathcal{P}\backslash\mathcal{P}_3$. The decoding orders achieving $\mathcal{R}_3(\mathcal{P}_3)$ are UE $i \overset{(1)}{\rightarrow} x_{B2} \overset{(2)}{\rightarrow} (x_{A1}, x_{A2})$ and UE $j \overset{(1)}{\rightarrow} x_{B1} \overset{(2)}{\rightarrow} x_{B2}$. Moreover, the decoding orders for $\mathcal{R}_4(\mathcal{P}_4)$ are UE $i \overset{(1)}{\rightarrow} x_{A2} \overset{(2)}{\rightarrow} x_{A1}$ and UE $j \overset{(1)}{\rightarrow} x_{B1} \overset{(2)}{\rightarrow} x_{A2} \overset{(3)}{\rightarrow} x_{B2}$.

Proof: Please refer to Appendix B. ∎

3) UE $j \overset{(1)}{\rightarrow} (x_{A2}, x_{B2})$ and UE $i \overset{(1)}{\rightarrow} (x_{A2}, x_{B2})$: Recall that decoding the interference signals first is only possible if the SIC condition is fulfilled. In this case, the achievable rate region is given in Proposition 3.

Proposition 3. When UE $j \overset{(1)}{\rightarrow} (x_{A2}, x_{B2})$ and UE $i \overset{(1)}{\rightarrow} (x_{A2}, x_{B2})$, the achievable rate region is given by $\mathcal{R}_5(\mathcal{P}_5) \bigcup \mathcal{R}_6(\mathcal{P}_6) \bigcup \mathcal{R}_7(\mathcal{P}_7)$, where

$$\mathcal{R}_5(\mathcal{P}_5) \triangleq \bigcup_{\mathbf{p}\in\mathcal{P}_5} \left\{ \mathbf{r} \left| \begin{array}{l} r_{i,s} \le C\left(\frac{p_{i,s}}{\alpha_i}\right),\ s\in\{1,2\} \\ r_{i,1}+r_{i,2} \le C\left(\frac{p_{i,1}+p_{i,2}}{\alpha_i}\right) \\ r_{j,1} \le C\left(\frac{p_{j,1}}{p_{i,2}+\alpha_j}\right) \\ r_{j,2} \le C\left(\frac{p_{j,2}}{p_{j,1}+p_{i,2}+\alpha_j}\right) \end{array} \right. \right\}$$

$$\mathcal{R}_6(\mathcal{P}_6) \triangleq \bigcup_{\mathbf{p}\in\mathcal{P}_6} \left\{ \mathbf{r} \left| \begin{array}{l} r_{i,1} \le C\left(\frac{p_{i,1}}{\alpha_i+p_{j,2}\Delta}\right) \\ r_{i,2} \le C\left(\frac{p_{i,2}}{p_{i,1}+p_{j,2}+\alpha_i}\right) \\ r_{j,1} \le C\left(\frac{p_{j,1}}{\alpha_j}\right) \\ r_{j,2} \le C\left(\frac{p_{j,2}}{p_{i,2}+p_{j,1}+\alpha_j}\right) \end{array} \right. \right\}$$

$$\mathcal{R}_7(\mathcal{P}_7) \triangleq \bigcup_{\mathbf{p}\in\mathcal{P}_7} \left\{ \mathbf{r} \left| \begin{array}{l} r_{i,1} \le C\left(\frac{p_{i,1}}{p_{j,2}+\alpha_i}\right) \\ r_{i,2} \le C\left(\frac{p_{i,2}}{p_{i,1}+p_{j,2}+\alpha_i}\right) \\ r_{j,s} \le C\left(\frac{p_{j,s}}{\alpha_j}\right),\ s=1,2 \\ r_{j,1}+r_{j,2} \le C\left(\frac{p_{j,1}+p_{j,2}}{\alpha_j}\right) \end{array} \right. \right\}$$

with $\mathcal{P}_5 \triangleq \{\mathbf{p}\in\mathcal{P} \mid p_{i,1} < p_{j,1} + \alpha_j - \alpha_i\}$, $\mathcal{P}_6 = \mathcal{P}_7 = \mathcal{P}\backslash\mathcal{P}_5$, and $\Delta \triangleq \mathbf{1}[p_{i,2} > p_{i,1} - p_{j,1} - \alpha_j + \alpha_i]$. The decoding orders achieving $\mathcal{R}_5(\mathcal{P}_5)$ are UE $i \overset{(1)}{\rightarrow} x_{B2} \overset{(2)}{\rightarrow} (x_{A1}, x_{A2})$ and UE $j \overset{(1)}{\rightarrow} x_{B2} \overset{(2)}{\rightarrow} x_{B1}$. Moreover, $\mathcal{R}_6(\mathcal{P}_6)$ is achieved by the decoding orders

$$\text{UE } i \overset{(1)}{\rightarrow} x_{A2} \overset{(2)}{\rightarrow} \begin{cases} x_{A1}, & \text{if } \Delta = 1, \\ x_{B2} \overset{(3)}{\rightarrow} x_{A1}, & \text{otherwise,} \end{cases}$$

and UE $j \overset{(1)}{\rightarrow} x_{B2} \overset{(2)}{\rightarrow} x_{B1}$. Finally, $\mathcal{R}_7(\mathcal{P}_7)$ is achieved by the decoding orders UE $i \overset{(1)}{\rightarrow} x_{A2} \overset{(2)}{\rightarrow} x_{A1}$ and UE $j \rightarrow x_{A2} \overset{(2)}{\rightarrow} (x_{B1}, x_{B2})$.

Proof: Please refer to Appendix C. ∎

Remark 3. Different from conventional NOMA, file splitting in the proposed cache-aided NOMA scheme enables joint decoding opportunities. For example, joint decoding of x_{B1} and x_{B2} at UE j is possible in $\mathcal{R}_1(\mathcal{P}_1)$, as the two signals are received by UE j over the same AWGN channel with noise variance $p_{i,2} + \alpha_j$. Therefore, UE j can choose the decoding order for these files without restriction. Similarly, joint decoding of x_{A1} and x_{A2} is possible at UE i in $\mathcal{R}_3(\mathcal{P}_3)$ and $\mathcal{R}_5(\mathcal{P}_5)$, respectively. We note that employing file splitting in conventional NOMA would not increase the achievable rates at UEs. However, if a portion of file is cached at one of the UEs, the achievable rates of the UEs can be increased by employing file splitting in the proposed cache-aided NOMA as this enables CIC.

Finally, combining the results in Propositions 1–3, the overall achievable rate region is $\mathcal{R} \triangleq \bigcup_{n=1}^{7} \mathcal{R}_n(\mathcal{P}_n)$. Note that $\mathcal{R}_n(\mathcal{P}_n)$ can be written in general form as

$$\mathcal{R}_n(\mathcal{P}_n) = \bigcup_{\mathbf{p}\in\mathcal{P}_n} \left\{ \mathbf{r} \left| \begin{array}{l} \text{C2: } r_{ks} \le C_{k,s}(\mathbf{p}), k\in\{i,j\}, s\in\{1,2\} \\ \text{C3: } r_{k1}+r_{k2} \le C_{k,1,2}(\mathbf{p}), k\in\{i,j\} \end{array} \right. \right\} \tag{8}$$

where $C_{k,s}$ and $C_{k,1,2}$ are the respective capacity bounds for decoding signal x_{fs}, $s \in \{1,2\}$, and signals $\{x_{f1}, x_{f2}\}$ at user $k \in \{i,j\}$.

B. Rate and Power Allocation for Fast Delivery

Let T be the time required to complete the delivery of the requested files. We have

$$T = \max_{k\in\{i,j\},\, s\in\{1,2\}} \frac{\beta_{k,s}}{r_{k,s}}, \tag{9}$$

for $\mathbf{r} \in \mathcal{R}$, where $\beta_{k,1} \triangleq (\bar{c}_f - c_{kf})V_f$ and $\beta_{k,2} \triangleq (1-\bar{c}_f)V_f$ for $(k,f) \in \{(i,A),(j,B)\}$ denote the effective volume of data to be delivered to user k. To avoid trivial results, we assume throughout this section that $\beta_{k,1} + \beta_{k,2} > 0$, $\forall k \in \{i,j\}$, i.e., each user requests some video data that is not cached[2]. Consequently, the delivery time optimization problem is formulated as

$$\text{P1:} \min_{\mathbf{r}\in\mathcal{R},\, \mathbf{p}\in\mathcal{P},\, T\ge 0} T \tag{10}$$
$$\text{s.t.} \quad \text{C4: } r_{ks}T \ge \beta_{k,s},\ k\in\{i,j\}, s\in\{1,2\},$$

where C4 ensures completion of file delivery at time T.

Problem P1 is generally nonconvex as the capacity functions in C2 and C3 in (8) are not jointly convex with respect to $\mathbf{r}$ and $\mathbf{p}$, and C4 is bilinear. However, the optimal solution of Problem P1 can be obtained by solving a sequence of convex problems as will be shown in the following. In particular, assume that the optimal solution lies in $\mathcal{R}_n$. For each feasible power allocation, the rate region $\mathcal{R}_n$, cf. (8), reduces to a polyhedron. Consequently, the optimal rate allocation, denoted as $\mathbf{r}^* \triangleq (r_{i,1}^*, r_{i,2}^*, r_{j,1}^*, r_{j,2}^*)$, can be obtained as the rate tuple on the dominate face[3] of $\mathcal{R}_n$ [17]. For example, for $n = 1$, we have $r_{i,1}^* = C_{i,1}$ and $r_{i,2}^* = C_{i,2}$ as the rates of UE i are only constrained by C2. In contrast, as the rates of UE j are constrained by both C2 and C3, we have $r_{j,1}^* = C\left(\frac{p_{j,1}^*}{p_{j,2}^*+p_{i,2}^*+\alpha_j}\right)$

[2]Otherwise, $p_{k,1} = p_{k,2} = 0$ and $r_{k,1} = r_{k,2} = 0$.

[3]For a polyhedron, any point that lies outside the dominant face is dominated elementwise by some point on the dominant face [17].

Algorithm 1 Bisection search for ρ_n^*.

1: **initialization**: Given LB, UB, and tolerance ϵ;
2: **repeat**
3: $\rho_n \leftarrow (LB+UB)/2$;
4: Solve the feasibility problem of (11) for ρ_n;
5: **if** (11) is infeasible **then**
6: $UB \leftarrow \rho_n$;
7: **else**
8: $LB \leftarrow \rho_n$;
9: **end if**
10: **until** $UB - LB < \epsilon$.

and $r_{j,2}^* = C\left(\frac{p_{j,2}^*}{p_{i,2}^*+\alpha_j}\right)$ for decoding order $j \overset{(1)}{\rightarrow} x_{B1} \overset{(2)}{\rightarrow} x_{B2}$ and $r_{j,1}^* = C\left(\frac{p_{j,1}^*}{p_{i,2}^*+\alpha_j}\right)$ and $r_{j,2}^* = C\left(\frac{p_{j,2}^*}{p_{i,2}^*+p_{j,1}^*+\alpha_j}\right)$ for decoding order $j \overset{(1)}{\rightarrow} x_{B2} \overset{(2)}{\rightarrow} x_{B1}$, and the optimal power allocation $\mathbf{p}^* = (p_{i,1}^*, p_{i,2}^*, p_{j,1}^*, p_{j,2}^*)$. In the same manner, the optimal rate allocation for all $n \in \{1, \ldots, 7\}$ can be obtained.

Substituting the optimal rate allocations and letting $\rho_n = 1/T$, P1 can be equivalently reformulated as $T^* = \min_{n\in\{1,\ldots,7\}} \rho_n^*$ with

$$\rho_n^* \triangleq \max_{\mathbf{p}\in\mathcal{P}_n,\, \rho_n \geq 0} \rho_n \tag{11}$$
$$\text{s.t.} \quad \text{C5: } r_{k,s}^*(\mathbf{p}) \geq \rho_n \beta_{k,s}, k \in \{i,j\}, s \in \{1,2\}.$$

The optimal value ρ_n^* can be found iteratively by employing Algorithm 1. In particular, in each iteration, the feasibility of problem (11) is checked for a given ρ_n, cf. line 4. For given ρ_n, we have $\rho_n^* \geq \rho_n$ if (11) is feasible, i.e., ρ_n is a lower bound on ρ_n^*, and $\rho_n^* \leq \rho_n$ otherwise, i.e., ρ_n is an upper bound on ρ_n^*. Hence, a bisection search can be applied to iteratively update the value of ρ_n until the gap between the lower and the upper bounds vanishes, whereby ρ_n^* is obtained. Moreover, efficient convex optimization algorithms can be employed [18] in line 4 of Algorithm 1. This is because although C5 is a linear fractional constraint of the form $\log_2\left(1+\frac{\mathbf{a}^T\mathbf{p}}{\mathbf{b}^T\mathbf{p}+1}\right) \geq c$ for $\mathbf{a}, \mathbf{b} \in \mathbb{R}_+^4$ and $c \in \mathbb{R}_+$, it can be transformed into an equivalent convex constraint of the form $(\mathbf{a} - (2^c-1)\mathbf{b})^T \mathbf{p} \geq 2^c - 1$ such that an equivalent convex formulation of problem (11) is obtained.

IV. Performance Evaluation

In this section, the performance of the proposed cache-aided NOMA is evaluated by simulation. Consider a cell of radius $R = 2$ km, where the BS is deployed at the center of the cell and the strong and the weak users, UE i and UE j, are uniformly distributed on discs of radii $R_i = 0.2$ km and $R_j = 0.6$ km, respectively. For modeling the wireless channel, the 3GPP path loss model ("Urban Macro NLOS" scenario) in [19] is adopted. The small-scale fading coefficients are independent and identically distributed (i.i.d.) Rayleigh random variables. The video files have size $V_A = V_B =$500 MBytes. Moreover, the system has a bandwidth of 5 MHz. The noise power spectral density is -172.6 dBm/Hz. Finally, we set the maximal transmit power as $P = 35$ dBm and the cache status as $c_{iA} = 0.2$, $c_{iB} = 0.8$, $c_{jA} = 0.8$, and $c_{jB} = 0.2$.

A. Baseline Schemes

1) Baseline 1 (Cache-aided orthogonal multiple access (OMA)): As baseline, we consider time-division multiple access (TDMA) for transmitting the uncached portions of the requested files. In particular, τ and $1-\tau$ fractions of time are allocated for transmission to UE i and UE j, respectively, where $\tau \in [0,1]$. Consequently, the capacity region for all possible time allocations is given by $\mathcal{R}_{\text{OMA}} = \bigcup_{\tau\in[0,1]}\left\{(r_i, r_j) \,\middle|\, r_i \leq \tau C(\frac{P}{\alpha_i}), r_j \leq (1-\tau)C(\frac{P}{\alpha_j})\right\}$. Note that, with Baseline 1, caching only facilitates conventional offloading of the hit cached data.

2) Baseline 2 (Conventional NOMA with and without caching): If caching is possible, Baseline 2 is a straightforward combination of caching and NOMA, whereby the requested data hit by the cache is offloaded and only the remaining data is transmitted by applying NOMA. If caching is not possible, Baseline 2 reduces to the conventional NOMA scheme. In both cases, the BS transmits signals $x = \sqrt{p_i}x_A + \sqrt{p_j}x_B$ for delivering files W_A and W_B, where the power allocations p_i and p_j satisfy $\mathcal{P}_{\text{NOMA}} \triangleq \left\{(p_i, p_j) \in \mathbb{R}_+^2 \mid p_i + p_j \leq P\right\}$. The received signals at UEs i and j are given by,

$$\begin{aligned} y_i &= h_i\left(\sqrt{p_i}x_A + \sqrt{p_j}x_B\right) + z_i, \\ y_j &= h_j\left(\sqrt{p_i}x_A + \sqrt{p_j}x_B\right) + z_j. \end{aligned} \tag{12}$$

For Baseline 2, the same capacity region $\mathcal{R}_{\text{NOMA}}(\mathcal{P}_{\text{NOMA}}) = \bigcup_{\mathbf{p}\in\mathcal{P}_{\text{NOMA}}}\left\{(r_i, r_j) \,\middle|\, r_i \leq C(\frac{p_i}{\alpha_i}), r_j \leq C(\frac{p_j}{p_i+\alpha_j})\right\}$ is achieved by SIC with and without caching [17], where x_B is decoded and canceled before decoding x_A at UE i. For Baselines 1 and 2, the rate and time/power allocation is optimized for minimization of the delivery time in a similar manner as for the proposed cache-aided NOMA scheme.

B. Simulation Results

In Fig. 3, we compare the achievable rate regions of the proposed cache-aided NOMA scheme and the baseline schemes for $\alpha_i = 10^{-3}$ and $\alpha_j = 10^{-2}$. For the proposed scheme, the rate achievable by UE k is given by $r_{k,1} + r_{k,2}$, $k \in \{i,j\}$. Note that the achievable rate regions of all considered schemes are independent of the values of c_{kf}, $(k,f) \in \{(i,A),(j,B)\}$. In particular, Baseline 2 with and without caching achieves the same rate region. From Fig. 3, we observe that all considered schemes achieve the same corner points $(0, 10.0)$ and $(13.2, 0)$, since the maximal rate for each UE is fundamentally limited by its channel status. Baseline 1 achieves the smallest rate region as it employs OMA to avoid interference. As NOMA introduces additional degrees of freedom for the users, Baseline 2 has a larger achievable rate region than Baseline 1. The expansion of the rate region is more significant for the weak user than for the strong user since the strong user consumes a small transmit power, and hence, causes little interference to the weak user. The proposed cache-aided NOMA scheme achieves the largest rate region among all considered schemes as joint CIC and SIC allows more interference to be canceled compared to Baseline 2 which can only perform SIC. This translates into a large sum rate gain for the proposed scheme. With the proposed scheme, significant performance gains are possible for both the weak and the strong user due to CIC.

In Fig. 4, we show the optimal average delivery times of the proposed cache-aided NOMA scheme and Baselines 1 and 2 as functions of the distance of the weak UE to the BS R_j. The performance is averaged over different realizations of the user locations and the channel fading. For a given R_j, as expected from the achievable rate region results in Fig. 3, Baseline 1 requires the longest time to complete video file delivery. The proposed cache-aided NOMA scheme outperforms both Baseline 2 without caching and Baseline 2 with caching. This is due to the exploitation of CIC, which

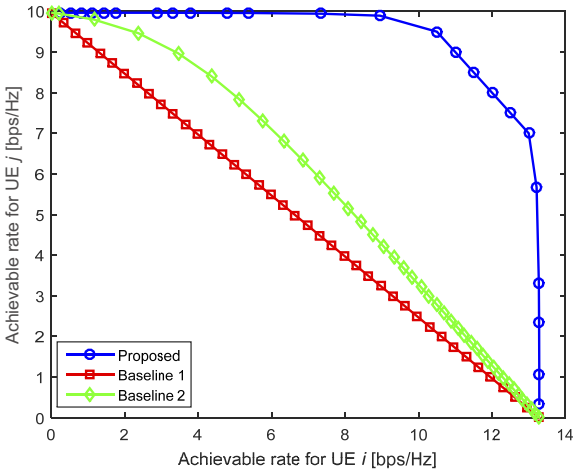

Fig. 3. Achievable rate region of the proposed scheme and Baselines 1 and 2 for $\alpha_i = 10^{-3}$ and $\alpha_j = 10^{-2}$.

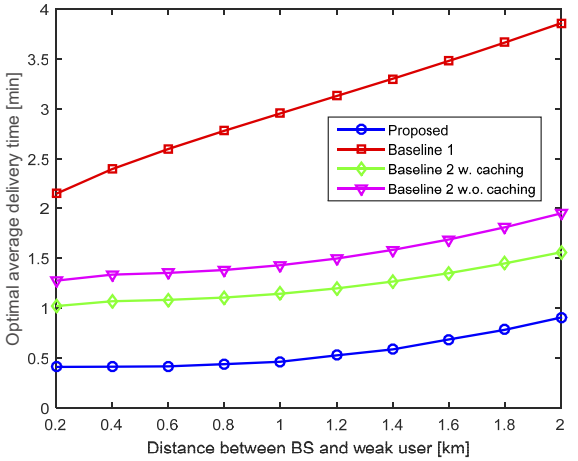

Fig. 4. Optimal average delivery time versus the distance between the weak user and the BS.

is possible only with the proposed joint caching and NOMA transmission design. However, different from the achievable rate region, the delivery times of Baseline 1, Baseline 2 with caching, and the proposed scheme critically depend on the amount of cached data. As R_j increases, UE j, the weak user, suffers from an increased path loss, which in turn reduces the channel gain of UE j. Moreover, since the delivery time of the weak user dominates the overall delivery time, we observe from Fig. 4 that the optimal delivery time increases with R_j for all considered schemes. However, Baseline 1 is the least efficient among the considered schemes, and its delivery time increases by about 80% as R_j increases from 0.2 km to 2 km. By exploiting NOMA and the resulting increased degrees of freedom, Baseline 2 effectively reduces the performance degradation caused by the weak user. For example, even without caching, the delivery time of Baseline 2 is 40% (50%) lower than that of Baseline 1 when UE j is located at $R_j = 0.2$ km ($R_j = 2$ km). Moreover, when a cache is available, Baseline 2 can also exploit caching for offloading of the delivery data, which further reduces the delivery time compared to Baseline 1 by an additional 12% (10%) for $R_j = 0.2$ km ($R_j = 2$ km). The proposed scheme enjoys the best performance and its delivery time is about 80% lower than that of Baseline 1 for the considered values of R_j.

V. Conclusion

In this paper, a joint caching and NOMA transmission design was presented for spectrally efficient downlink communication. The proposed scheme exploits unrequested cached data for cancellation of NOMA interference, which is not possible with separate caching and NOMA transmission. The achievable rate region of the proposed cache-aided NOMA scheme was characterized, and the optimal decoding order and the optimal power and rate allocations for minimization of the delivery time were investigated. Simulation results showed that the proposed scheme can significantly expand the achievable downlink rate region for both the strong and the weak users. Moreover, the delivery time of both users can be effectively reduced to achieve fast video delivery. For ease of illustration, the proposed cache-aided NOMA was only evaluated for the important case of two paired NOMA users. The extension of cache-aided NOMA to multiple users will be considered in future work.

Appendix A
Proof of Proposition 1

As $i \overset{(1)}{\rightarrow} x_{A1}$, $r_{i,1} \leq C_{i,1}$ is achievable for decoding x_{A1} at UE i. To derive the achievable rate region, we need to check the decodability of the interfering signals x_{B2} and x_{A2} at UE i and j, respectively. Let us consider the following two power regions.

(1) For $\mathbf{p} \in \mathcal{P}_1 \backslash \mathcal{P}_2$, we have $C\left(\frac{p_{i,2}}{p_{j,1}+\alpha_j}\right) < C\left(\frac{p_{i,2}}{p_{j,2}+\alpha_i}\right)$, i.e., UE j cannot decode x_{A2} before decoding x_{B1} as the SIC decoding condition is not met. Also, for any $(p_{j,1}, p_{j,2}) \in \mathbb{R}_+^2$, x_{A2} cannot be decoded before decoding x_{B2} at UE j as

$$C\left(\frac{p_{i,2}}{p_{j,1}+p_{j,2}+\alpha_j}\right) < C\left(\frac{p_{i,2}}{p_{j,2}+\alpha_j}\right) < C\left(\frac{p_{i,2}}{p_{j,2}+\alpha_i}\right). \tag{13}$$

On the other hand, for any $(p_{j,1}, p_{j,2}) \in \mathbb{R}_+^2$, x_{B2} can be always decoded and canceled at UE i before x_{A2} is decoded as $\alpha_i < \alpha_j$; and hence, $r_{i,2} \leq C_{i,2}$ is achievable. In contrast, UE j cannot decode x_{A2} in any case. Consequently, the feasible decoding orders are UE $i \overset{(2)}{\rightarrow} x_{B2} \overset{(3)}{\rightarrow} x_{A2}$ and UE $j \overset{(1)}{\rightarrow} (x_{B1}, x_{B2})$, whereby rate region $\mathcal{R}_1(\mathcal{P}_1 \backslash \mathcal{P}_2)$ is achieved.

(2) For $\mathbf{p} \in \mathcal{P}_2$, we have $C\left(\frac{p_{i,2}}{\alpha_i}\right) > C\left(\frac{p_{i,2}}{p_{j,1}+\alpha_j}\right) > C\left(\frac{p_{i,2}}{p_{j,2}+\alpha_i}\right)$, i.e., x_{A2} can be decoded at UE j before x_{B1} is decoded if and only if UE $i \overset{(2)}{\rightarrow} x_{A2}$. Assume x_{A2} is decoded last at UE i such that UE j cannot decode x_{A2} in any case. Then, rate region $\mathcal{R}_1(\mathcal{P}_2)$ is achievable. On the other hand, suppose x_{A2} is decoded first at UE i. Then, UE j can achieve a higher rate for $r_{j,1}$ by decoding x_{A2} before decoding x_{B1}, which is only possible after x_{B2} has been decoded according to (13). Thus, the rate region $\mathcal{R}_2(\mathcal{P}_2)$ is achievable.

Therefore, the rate region $\mathcal{R}_1(\mathcal{P}_1) \bigcup \mathcal{R}_2(\mathcal{P}_2)$ is achievable, and any rate vector outside the region $\mathcal{R}_1(\mathcal{P}_1) \bigcup \mathcal{R}_2(\mathcal{P}_2)$ cannot be achieved by SIC decoding. This completes the proof.

Appendix B
Proof of Proposition 2

By decoding x_{B1} first, $r_{j,1} \leq C\left(\frac{p_{j,1}}{p_{i,2}+p_{j,2}+\alpha_j}\right)$ is achievable for UE j. To obtain the achievable rate region, two power regions have to be considered.

(1) For $\mathbf{p} \in \mathcal{P}_3$, we have

$$C\left(\frac{p_{i,2}}{p_{i,1}+p_{j,2}+\alpha_i}\right) > C\left(\frac{p_{i,2}}{p_{j,2}+\alpha_j}\right), \quad (14)$$

$$C\left(\frac{p_{j,2}}{p_{i,1}+p_{i,2}+\alpha_i}\right) > C\left(\frac{p_{j,2}}{p_{i,2}+\alpha_j}\right), \quad (15)$$

which imply that x_{A2} cannot be decoded at UE j in general, cf. (14), but x_{B2} can always be decoded at UE i, cf. (15). Consequently, UE j can decode x_{B2} only by treating x_{A2} as noise whereas UE i will first decode x_{B2} and cancel its contribution to the received signal before decoding x_{A1} and x_{A2}. Therefore, the achievable rate region is given by $\mathcal{R}_3(\mathcal{P}_3)$.

(2) For $\mathbf{p} \in \mathcal{P}_4$, we have

$$C\left(\frac{p_{i,2}}{p_{i,1}+p_{j,2}+\alpha_i}\right) < C\left(\frac{p_{i,2}}{p_{j,2}+\alpha_j}\right), \quad (16)$$

$$C\left(\frac{p_{j,2}}{p_{i,1}+p_{i,2}+\alpha_i}\right) < C\left(\frac{p_{j,2}}{p_{i,2}+\alpha_j}\right), \quad (17)$$

$$C\left(\frac{p_{j,2}}{p_{i,1}+\alpha_i}\right) < C\left(\frac{p_{j,2}}{\alpha_j}\right). \quad (18)$$

That is, at UE i, x_{B2} cannot be decoded first, cf. (17). Hence, we only need to consider UE $i \overset{(1)}{\to} x_{A2}$. In this case, UE j is able to cancel the interference from x_{A2} before decoding x_{B2} due to (16). However, at UE i, x_{B2} cannot be canceled before decoding x_{A1} due to (18), i.e., UE $i \overset{(2)}{\to} x_{B2}$ is infeasible. Therefore, the achievable rate region is given by $\mathcal{R}_4(\mathcal{P}_4)$, where UE i cannot decode x_{B2} while UE j can decode and cancel x_{A1} before decoding x_{B2}. Therefore, the rate region in Proposition 2 is achievable, which completes the proof.

Appendix C
Proof of Proposition 3

First, assume UE $j \overset{(1)}{\to} x_{B2}$ and UE $i \overset{(1)}{\to} (x_{A2}, x_{B2})$. If $\mathbf{p} \in \mathcal{P}_5$, we have

$$C\left(\frac{p_{j,2}}{p_{i,1}+p_{i,2}+\alpha_i}\right) > C\left(\frac{p_{j,2}}{p_{i,2}+p_{j,1}+\alpha_j}\right), \quad (19)$$

$$C\left(\frac{p_{i,2}}{\alpha_i}\right) > C\left(\frac{p_{i,2}}{p_{i,1}+\alpha_i}\right) > C\left(\frac{p_{i,2}}{p_{j,1}+\alpha_j}\right). \quad (20)$$

By (19), UE i can decode and cancel x_{B2} as $\alpha_i < \alpha_j$. Thus, UE $i \overset{(1)}{\to} x_{B2}$, which leads to the residual received signals $y_i^{(1)} = h_i\left(\sqrt{p_{i,1}}x_{A1} + \sqrt{p_{i,2}}x_{A2}\right) + z_i$ and $y_j^{(1)} = h_j\left(\sqrt{p_{j,1}}x_{B1} + \sqrt{p_{i,2}}x_{A2}\right) + z_j$. By (20), UE j cannot decode x_{A2} based on $y_j^{(1)}$. Therefore, the decoding orders UE $i \overset{(1)}{\to} x_{B2} \overset{(2)}{\to} (x_{A1}, x_{A2})$ and UE $j \overset{(1)}{\to} x_{B2} \overset{(2)}{\to} x_{B1}$ are feasible and achieve rate region $\mathcal{R}_5(\mathcal{P}_5)$.

However, if $\mathbf{p} \in \mathcal{P}_6$, UE i cannot decode x_{B2} first due to (19). Then, for the assumption of UE $i \overset{(1)}{\to} (x_{A2}, x_{B2})$, we only need to consider the case UE $i \overset{(1)}{\to} x_{A2}$. We have

$$C\left(\frac{p_{i,2}}{p_{j,1}+\alpha_j}\right) > C\left(\frac{p_{i,2}}{p_{i,1}+\alpha_i}\right), \quad (21)$$

i.e., UE j can cancel x_{A2} before decoding x_{B1}. On the other hand, UE i cannot cancel x_{B2} before decoding x_{A1} unless $\Delta = 0$, whereby we have $C\left(\frac{p_{j,2}}{p_{i,2}+p_{j,1}+\alpha_j}\right) < C\left(\frac{p_{j,2}}{p_{i,1}+\alpha_i}\right)$. Hence, rate region $\mathcal{R}_6(\mathcal{P}_6)$ is achievable.

Next, assume UE $j \overset{(1)}{\to} x_{A2}$ and UE $i \overset{(1)}{\to} x_{A2}$, which requires $C\left(\frac{p_{i,2}}{p_{j,1}+p_{j,2}+\alpha_j}\right) > C\left(\frac{p_{i,2}}{p_{i,1}+p_{j,2}+\alpha_i}\right)$, or equivalently, $\mathbf{p} \in \mathcal{P}_7$. In this case, as $C\left(\frac{p_{j,2}}{\alpha_i}\right) > C\left(\frac{p_{j,2}}{p_{j,1}+\alpha_j}\right) > C\left(\frac{p_{j,2}}{p_{i,1}+\alpha_i}\right)$, UE i cannot decode x_{B2}. Hence, the decoding orders UE $i \overset{(1)}{\to} x_{A2} \overset{(2)}{\to} x_{A1}$ and UE $j \overset{(1)}{\to} x_{A2} \overset{(2)}{\to} (x_{B1}, x_{B2})$ are feasible and achieve rate region $\mathcal{R}_7(\mathcal{P}_7)$.

Finally, for UE $j \overset{(1)}{\to} x_{A2}$ and UE $i \overset{(1)}{\to} x_{B2}$, feasible power and rate allocations do not exist. In particular, for such a rate region to exist, the following inequalities would have to hold,

$$C\left(\frac{p_{i,2}}{p_{j,1}+p_{j,2}+\alpha_j}\right) > C\left(\frac{p_{i,2}}{p_{i,1}+\alpha_i}\right), \quad (22)$$

$$C\left(\frac{p_{j,2}}{p_{i,1}+p_{i,2}+\alpha_i}\right) > C\left(\frac{p_{j,2}}{p_{j,1}+\alpha_j}\right), \quad (23)$$

which ensure feasibility of UE $j \overset{(1)}{\to} x_{A2}$ and UE $i \overset{(1)}{\to} x_{B2}$, respectively. Eqs. (22) and (23) are equivalent to $p_{i,1} - p_{j,1} > p_{j,2} + \alpha_j - \alpha_i$ and $p_{i,1} - p_{j,1} < \alpha_j - \alpha_i - p_{j,2}$, respectively, which lead to $p_{i,2} + p_{j,2} < 0$. That is, (22) and (23) cannot be met for feasible powers. Therefore, the rate region in Proposition 3 is achievable, which completes the proof.

References

[1] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. Leung, "Cache in the air: Exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 131–139, Feb. 2014.

[2] V. W. S. Wong, R. Schober, D. W. K. Ng, and L.-C. Wang, *Key Technologies for 5G Wireless Systems*. Cambridge University Press, 2017.

[3] X. Ge, S. Tu, M. Guoqiang, C.-X. Wang, and T. Han, "5G ultra-dense cellular networks," *IEEE Wireless Commun.*, vol. 23, no. 1, pp. 72–79, Feb. 2016.

[4] L. Xiang, D. W. K. Ng, T. Islam, R. Schober, V. W. S. Wong, and J. Wang, "Cross-layer optimization of fast video delivery in cache- and buffer-enabled relaying networks," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 11 366–11 382, Dec. 2017.

[5] L. Xiang, D. W. K. Ng, R. Schober, and V. W. Wong, "Cache-enabled physical-layer security for video streaming in backhaul-limited cellular networks," to be published in *IEEE Trans. Wireless Commun.*.

[6] ——, "Secure video streaming in heterogeneous small cell networks with untrusted cache helpers," to be published in *IEEE Trans. Wireless Commun.*.

[7] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.

[8] ——, "Decentralized coded caching attains order-optimal memory-rate tradeoff," *IEEE/ACM Trans. Netw.*, vol. 23, no. 4, pp. 1029–1040, Aug. 2015.

[9] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE Veh. Technology Conf. (VTC Spring)*, Dresden, Germany, Jun. 2013, pp. 1–5.

[10] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas Commun.*, vol. 35, pp. 2181–2195, Oct. 2017.

[11] 3GPP TR 36.859, "Study on downlink multiuser superposition transmission (MUST) for LTE (Release 13)," Dec. 2015.

[12] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.

[13] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 1077–1091, Mar. 2017.

[14] Z. Ding, R. Schober, and H. V. Poor, "A general MIMO framework for NOMA downlink and uplink transmission based on signal alignment," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4438–4454, Aug. 2016.

[15] Z. Ding, P. Fan, G. K. Karagiannidis, R. Schober, and H. V. Poor, "NOMA assisted wireless caching: Strategies and performance analysis," *arXiv preprint arXiv:1709.06951*, 2017.

[16] J.-P. Hong and W. Choi, "User prefix caching for average playback delay reduction in wireless video streaming," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 377–388, Jan. 2016.

[17] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.

[18] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.

[19] 3GPP TR 36.814, "Further advancements for E-UTRA physical layer aspects (Release 9)," Mar. 2010.