Linear Network Coded Wireless Caching in Cloud Radio Access Network

Long Shi, *Member, IEEE*, Kui Cai, *Senior Member, IEEE*, Tao Yang, *Member, IEEE*, Taotao Wang, *Member, IEEE*, and Jun Li, *Senior Member, IEEE*

***Abstract*—This paper investigates a cache-aided cloud radio access network (C-RAN), comprising a central unit, K base stations (BSs) each with N_T antennas, and M users each with N_R antennas, where each BS and user have local caches to store some popular contents from the central unit. For this cache-aided network, we propose the linear network coded (NC) wireless caching that consists of linear wireless network coding assisted cache placement phase and signal-space alignment (SSA) enabled content delivery phase. In the cache placement phase, we design a joint NC caching function at the BSs to store linear combinations of messages from the central unit, as a form of linear wireless network coding. In the content delivery phase, we design the SSA pattern based on the NC caching to guide the precoding designs at BSs. Then, each user can reliably decode its requested messages by receiver shaping and reverse NC operation. The primary contribution of this work is to achieve the coding gain induced by the integration of linear wireless network coding and SSA, which has been not exploited in the field of wireless coded caching. In particular, to deal with high temporal variability of user requests, we show that the proposed cache placement is invariant to different user requests in the worst-case caching, without any shared caching messages at different BSs. Furthermore, we verify that the proposed scheme is also compatible with the insufficient caching scenario at the BSs. In addition, we analyze the achievable sum degrees of freedom (DoF) for the proposed caching network. Both analytical and numerical results verify that the proposed caching scheme achieves a higher sum DoF than the existing related works.**

***Index Terms*—Wireless coded caching, linear wireless network coding, signal-space alignment, degrees of freedom**

I. Introduction

In modern wireless networks, fronthaul links are threatened by an alarming "digestive disease", due to the huge congestion caused by explosive growth of wireless traffic. Even worse, the ever-growing traffic congestion will soon reach the capacity limitation in the today's network generation. Therefore, how to alleviate the fronthaul congestion while meeting the peak traffic demands is a matter of great urgency [1]–[3].

Recent research unveils that multimedia delivery is a driving factor of the wireless traffic, of which duplicate downloads of a few popular contents (e.g., music or videos) occupy a significant portion [2], [3]. This finding drives us to reduce the redundant delivery through the fronthaul to alleviate the traffic congestion. To deal with this challenge, caching revives and come into play in wireless networks. Following the spirit of web caching, one way to reduce the wireless traffic burden is to employ memories distributed across the networks. Leveraging these memory units, wireless edges (such as base station (BS) and mobile device) pre-download and store the content via the fronthaul during off-peak hours (cache placement phase). Then, the fronthaul burden can be reduced, since the users can download these popular contents directly from its local cache or the nearby BSs in peak hours (content delivery phase), rather than from the fronthaul. Recently, wireless caching has been applied to a wide ranges of wireless networks [4]–[9]. Standing apart from web caching, wireless caching, operated in the physical layer, attains the significant gains integrated with advanced physical-layer coding technologies [7].

The seminal work [10] proposed coded caching to investigate fundamental limits of cache-aided broadcasting networks (consisting of one transmitter and multiple receivers), assuming a wired, shared, and error-free link. The bit-wise XOR network coding was employed in the delivery phase. Superior to uncoded caching, coded caching can explore the global caching gain from the coded multicasting transmission, in addition to the local caching gain [10], [11]. As follow-up, [12]–[14] proposed various coded caching strategies for wireless broadcast channels. Coded caching also manifests its benefits in wireless radio access networks (RANs) [8], [15]–[19]. The major challenges in the cache-aided RAN lie in the cache placement at transmitters and receivers and the interference management induced by different user requests in the content delivery. For example, [8], [15] optimized the beamforming vectors to minimize the sum of backhaul and power consumption cost in the cache-enabled cloud RANs (C-RANs), and [16], [17] studied the achievable degrees-of-freedom (DoF) under arbitrary cache size at both single-antenna transmitters and receivers. Considering multiple antennas at the wireless interference networks, [18], [19] adopted the bit-wise XOR network coding and interference management techniques such as interference alignment or interference neutralization, as different cooperation strategies among the transmitters are allowed in the cache placement. The ultimate

The work of L. Shi and K. Cai was supported by Singapore Ministry of Education Academic Research Fund Tier 2 MOE2016-T2-2-054. The work of T. Yang was supported by Beihang "Zhuoyue" Program ZG216s2066. The work of J. Li was supported in part by National Key R&D Program under Grants 2018YFB1004800, in part by National Natural Science Foundation of China under Grants 61727802 and 61872184. Part of this work was presented at the IEEE ICC Workshop, Kansas City, MO, USA in May 2018. (*Corresponding author: Kui Cai.*)

L. Shi and J. Li are with School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China (email: slong1007@gmail.com; jun.li@njust.edu.cn). K. Cai is with Science and Math Cluster, Singapore University of Technology and Design, Singapore (email: cai_kui@sutd.edu.sg). T. Yang is with School of Electronics and Information Engineering, Beihang University, Beijing, China (email: tyang@buaa.edu.cn). T. Wang is with the College of Information Engineering, Shenzhen University, Shenzhen, China (e-mail: ttwang@szu.edu.cn).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2020.3035607, IEEE Transactions on Communications

goal of those works is to investigate the maximum DoF for the wireless coded caching networks. Later on, [14] also studied the symmetric rate in the finite signal-to-noise ratio (SNR) regimes. Under fronthaul rate consideration, uncoded caching scheme in [34] characterized the normalized delivery time for the cloud and cache-aided RAN under different fronthaul-edge transmission modes, where the zero-forcing precoding is employed in the delivery phase.

In this paper, we address a series of issues that are not considered in the existing works on wireless coded caching. *First*, the bit-wise XOR network coding is not the optimal solution to accommodate the fading and interference, even in the wireless networks without cache. Recent works on the linear wireless network coding shows an elegant way of harnessing the interference to improve the transmission reliability while retaining the maximum network throughput [20]–[24]. The principle of linear wireless network coding is to compute the linear combinations of user messages according to wireless fading and the structure of interference. The goal of this paper is to design the linear wireless network coding in the wireless caching by exploiting the characteristics of wireless channels and interference. *Second*, the related works have not explored the extra coding gain brought by the nature of wireless network coding, as the interference mitigation in the delivery phase mainly relies on the shared cache placement among different transmitters rather than the structure of wireless network coding. Targeting at the coding gain, interference management in the delivery phase catering to the wireless network coding operated caching remains open. *Third*, wireless caching is featured as high temporal variability of the user requests. The frequent variation of caching placement at BSs will result in high latency and cost. Therefore, an invariant cache placement compatible with different user requests is preferable. Inspired by the existing works in [5], [6], [10], [11], [18], [19], the goal of this paper is to design the linear wireless network coding aided cache placement that is invariant to different user requests.

To cope with these problems, we proposed the linear wireless network coding operated wireless caching, referred to as *linear network coded (NC) wireless caching*. To illustrate the proposed caching strategy, we consider a cache-aided C-RAN that consists of a central unit, K BSs, and M users. Each BS and user are equipped with local caches and multiple antennas. In the cache placement phase, we design a joint NC caching function at BSs to prefetch multiple linear combinations of messages from the central unit, as a form of linear wireless network coding. In the content delivery phase, we propose a joint precoding design at all BSs to perform the signal space alignment (SSA) through the bin matrix designs at users. By receiver shaping and reverse NC operation, each user can successfully retrieve its requested messages. The key contributions of this paper are summarized as follows:

- We design the linear NC wireless caching in the cache-aided C-RAN. The core designs of this network lie in the linear NC assisted cache placement phase and SSA enabled content delivery phase (see Sections III and IV). In particular, each BS stores the NC caching messages as the linear combinations of messages in the central unit, without any shared caching messages at different BSs (see Section III-A). To guide the precoding design, we design the bin matrix to characterize the SSA patterns under different user requests (see Sections IV-A and IV-B).
- We characterize the achievable sum DoF for the proposed caching scheme. Because of the integration of linear wireless network coding and SSA, the proposed scheme attains the extra coding gain, contributing to a higher sum DoF than the existing related works (see Section V).
- We verify that the proposed cache placement is invariant to different user requests in the worst-case caching (i.e., all user requests are distinct) through the pre-designed NC caching matrix at the BSs. In addition, we show that the proposed wireless caching scheme is also applicable in the insufficient caching at some BSs under the rate constraint fronthaul (see Sections VI).

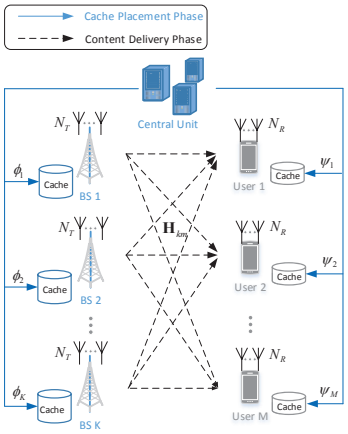


Fig. 1. System model of the cache-aided C-RAN with a central unit, K BSs, and M users.

II. System Model

Fig. 1 shows a cache-aided C-RAN that consists of a central unit, K BSs, and M users. Each BS has N_T antennas, and each user is equipped with N_R antennas. The content in the central unit consists of M message vectors, denoted by $\mathbf{w}_{1\sim M} = [\mathbf{w}_1^{\mathcal{T}}, \mathbf{w}_2^{\mathcal{T}}, \ldots, \mathbf{w}_M^{\mathcal{T}}]^{\mathcal{T}}$, where each subvector $\mathbf{w}_m$ contains multiple messages. All BSs are connected to the central unit through the error-free fronthaul layer in the centralized manner. The local caches of finite size are equipped at BSs and users respectively. As depicted in Fig. 1, the wireless caching is composed of cache placement phase and content delivery phase.

Cache placement phase: All BSs and users have the access to the entire content in the central unit and prefetch some popular messages at their local caches with finite sizes, according to pre-assigned caching functions. Define $\mathbf{u}_k = \phi_k(\mathbf{w}_{1\sim M})$ and $\mathbf{v}_m = \psi_m(\mathbf{w}_{1\sim M})$ as the caching message vectors associated with the caching functions ϕ_k and ψ_m at BS k and user m, $k = 1, 2, \ldots, K$ and $m = 1, 2, \ldots, M$, respectively. Note that the caching function ϕ_k at BS k is independent of the user requests, since the BS is unaware of future requests of the users in the content placement phase.

Content delivery phase: Each user m requests a message vector $\mathbf{w}_{\gamma_m}$ from the central unit, $\gamma_m \in \{1, 2, \ldots, M\}$. Let

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2020.3035607, IEEE Transactions on Communications

$\boldsymbol{\gamma} = [\gamma_1\ \gamma_2\ \ldots\ \gamma_M]$ denote the request vector collected from all users, where γ_m corresponds to user m's request. In this phase, all BSs are informed of these requests and proceed by transmitting a function of caching messages over wireless channels (i.e., air-interface layer). This phase is equivalent to a cache-aided downlink distributed multiple-input multiple-output (MIMO) transmission.

Each BS k first processes its caching message vector to be a space-time signal block $\mathbf{X}_k$. All BSs broadcast their respective signal blocks to all users simultaneously. This paper considers a total power constraint $\sum_{k=1}^{K} E\{\mathrm{tr}(\mathbf{X}_k^{\mathcal{H}}\mathbf{X}_k)\} \leq P_T$, where $E\{\cdot\}$ denotes the expectation and $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. Then, user m receives

$$\mathbf{Y}_m = \sum_{k=1}^{K} \mathbf{H}_{k,m}\mathbf{X}_k + \mathbf{Z}_m,\ m = 1, 2, \ldots, M, \tag{1}$$

where $\mathbf{H}_{k,m}$ denotes the channel coefficient matrix between BS k and user m, and $\mathbf{Z}_m$ denotes the AWGN matrix at user m with i.i.d. entries of zero mean and variance σ_z^2. This paper considers that the channel coefficients remain invariant over the duration of each block transmission and vary over different blocks. Suppose that the channel information is available at all BSs and users. The SNR per BS is defined as $\rho = \frac{P_T}{K\sigma_z^2}$.

Upon receiving $\mathbf{Y}_m$ in (1), user m decodes its requested messages $\mathbf{w}_{\gamma_m}$. Let $\hat{\mathbf{w}}_{\gamma_m}$ denote the decoded message vector of user m. The decoding error probability is defined as

$$\mathrm{Pr}_e = \mathrm{Pr}\big\{[\mathbf{w}_{\gamma_1}, \mathbf{w}_{\gamma_2}, \ldots, \mathbf{w}_{\gamma_M}] \neq [\hat{\mathbf{w}}_{\gamma_1}, \hat{\mathbf{w}}_{\gamma_2}, \ldots, \hat{\mathbf{w}}_{\gamma_M}]\big\}. \tag{2}$$

DoF metric: Let R_m denote the transmission rate of user m's messages. In the wireless caching network, an information rate tuple $(R_1, R_2, \ldots, R_M)$ is said to be achievable, if there exists a channel code $(R_1, R_2, \ldots, R_M, n, \mathrm{Pr}_e)$ such that $\mathrm{Pr}_e \leq \epsilon$ for any $\epsilon > 0$ as the coding block length n goes to infinity. Then, a DoF tuple $(d_1, d_2, \ldots, d_M)$ is said to be achievable if the rate of each user scales as $\lim\limits_{\rho\to\infty} \frac{R_m}{\log \rho} = d_m$ [25]. Furthermore, an achievable sum DoF is $\underline{d} = \sum_{m=1}^{M} d_m$.

III. Proposed Linear NC Wireless Caching

In this section, we propose a $K \times M$ linear network coded (NC) wireless caching network, as shown in Fig. 2. We first consider the worst-case caching scenario, where all user requests are distinct and no local cache is deployed at each user. That is, $\gamma_m \neq \gamma_{m'}$, $\forall m, m' \in \{1, \ldots, M\}$ and $m \neq m'$. Furthermore, the caching at both BSs and users will be elaborated in Section VI.

A. Linear Wireless Network Coding Assisted Cache Placement Phase

In practice, the user request varies independently and frequently over different block transmissions and time slots. Driven by the existing works in [5], [6], [10], [11], [18], [19], we propose a unified design of the joint NC caching function for K BSs to fulfill different user requests.

Let $\mathbf{w}_m = [w_{m,1}\ w_{m,2} \ldots\ w_{m,L_m^{\mathrm{o}}}]^{\mathcal{T}}$ be the mth message vector in the central unit, $m = 1, 2, \ldots, M$. Consider that each

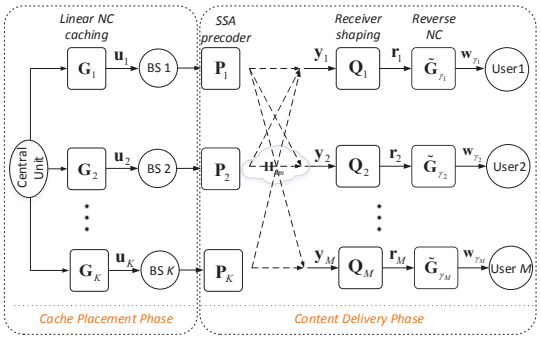


Fig. 2. System model of the proposed $K \times M$ linear NC wireless caching.

element of $\mathbf{w}_m$ is drawn i.i.d. from a finite field $\mathbb{F}_q$, i.e., $\mathbf{w}_m \in \mathbb{F}_q^{L_m^{\mathrm{o}}}$, where q corresponds to the modulation cardinality. Let $L^{\mathrm{o}} = L_1^{\mathrm{o}} + L_2^{\mathrm{o}} + \ldots + L_M^{\mathrm{o}}$ denote the total number of messages. In the cache placement phase, we design the linear wireless network coding assisted caching functions ϕ_k at BS k to store a length-L_k^{b} caching message vector as

$$\mathbf{u}_k = \phi_k(\mathbf{w}_{1\sim M}) = \mathbf{G}_k \otimes \mathbf{w}_{1\sim M},\ k = 1, 2, \ldots, K, \tag{3}$$

where $\mathbf{G}_k \in \mathbb{F}_q^{L_k^{\mathrm{b}} \times L^{\mathrm{o}}}$ is the NC caching matrix of BS k. Let $\mathbf{g}_{k,l_k}$ denote the l_kth row vector of $\mathbf{G}_k$ and $g_{k,l_k}[j]$ denote the (l_k, j)th element in $\mathbf{G}_k$, respectively. Then, $g_{k,l_k}[j] \in \mathbb{F}_q$, $l_k = 1, 2, \ldots, L_k^{\mathrm{b}}$ and $j = 1, 2, \ldots, L^{\mathrm{o}}$. The operation $\otimes$ denotes the multiplication in $\mathbb{F}_q$, i.e., $a \otimes b = ab \pmod q$.

We refer to $\mathbf{u}_k \in \mathbb{F}_q^{L_k^{\mathrm{b}}}$ as the NC caching message vector stored at BS k. Let $\mathbf{u}_k = [u_{k,1}\ u_{k,2}\ \ldots\ u_{k,L_k^{\mathrm{b}}}]^{\mathcal{T}}$ with $u_{k,l_k} \in \mathbb{F}_q$ being the ith NC caching message of $\mathbf{u}_k$, given by

$$\begin{aligned} u_{k,l_k} &= \mathbf{g}_{k,l_k} \otimes \mathbf{w}_{1\sim M} \\ &= g_{k,l_k}[1] \otimes w_{1,1} \ldots \oplus g_{k,l_k}[L_1^{\mathrm{o}}] \otimes w_{1,L_1^{\mathrm{o}}} \ldots \\ &\oplus g_{k,l_k}[L^{\mathrm{o}} - L_M^{\mathrm{o}} + 1] \otimes w_{M,1} \ldots \oplus g_{k,l_k}[L^{\mathrm{o}}] \otimes w_{M,L_M^{\mathrm{o}}}, \end{aligned} \tag{4}$$

for $l_k = 1, 2, \ldots, L_k^{\mathrm{b}}$. The operation $\oplus$ denotes the addition in $\mathbb{F}_q$, i.e., $a \oplus b = a + b \pmod q$. From (3) and (4), BS k prefetches L_k^{b} linear combinations of messages from the central unit rather than the messages themselves. Let $L^{\mathrm{b}} = L_1^{\mathrm{b}} + L_2^{\mathrm{b}} + \ldots + L_K^{\mathrm{b}}$ denote the total number of NC caching messages stored at K BSs. Throughout this paper, the total memory of all the BSs (i.e., L^b) is less than or equal to the total number of messages in the central unit (i.e., L^o).

The NC caching message vectors in (3) and (4) can be collectively expressed as

$$\mathbf{u}_{1\sim K} = [\mathbf{u}_1^{\mathcal{T}}\ \mathbf{u}_2^{\mathcal{T}}\ \ldots\ \mathbf{u}_K^{\mathcal{T}}]^{\mathcal{T}} = \mathbf{G}_{1\sim K} \otimes \mathbf{w}_{1\sim M}, \tag{5}$$

where $\mathbf{G}_{1\sim K} = [\mathbf{G}_1^{\mathcal{T}}\ \mathbf{G}_2^{\mathcal{T}}\ \ldots\ \mathbf{G}_K^{\mathcal{T}}]^{\mathcal{T}} \in \mathbb{F}_q^{L^{\mathrm{b}} \times L^{\mathrm{o}}}$ is referred to as the joint NC caching matrix. This paper considers that $\mathbf{G}_{1\sim K}$ has the full rank in $\mathbb{F}_q$. Note that $\mathbf{w}_{1\sim M}$ cannot be recovered from $\mathbf{u}_{1\sim K}$ if $\mathbf{G}_{1\sim K}$ is rank deficient over $\mathbb{F}_q$. Furthermore, it suffices to consider a full-rank $\mathbf{G}_{1\sim K}$ over $\mathbb{F}_q$ as a square matrix (i.e., $L^{\mathrm{b}} = L^{\mathrm{o}}$). For $L^{\mathrm{b}} > L^{\mathrm{o}}$, there exist some caching messages that are linearly dependent in $\mathbb{F}_q$, resulting in the waste of limited caching space at BSs.

Consider that $L = L^{\mathrm{b}} = L^{\mathrm{o}}$. In this case, L_k^b is simplified to L_k. Define $\tilde{\mathbf{G}}_{1\sim K} \in \mathbb{F}_q^{L\times L}$ as the *inverse NC matrix* of $\mathbf{G}_{1\sim K}$ in $\mathbb{F}_q$ such that $\tilde{\mathbf{G}}_{1\sim K} \otimes \mathbf{G}_{1\sim K} = \mathbf{I}_L$,

where $\mathbf{I}_L$ is an identity matrix with the size of $L \times L$. Let $\tilde{\mathbf{G}}_{1\sim K} = [\tilde{\mathbf{G}}_1^{\mathcal{T}}\ \tilde{\mathbf{G}}_2^{\mathcal{T}}\ \ldots\ \tilde{\mathbf{G}}_K^{\mathcal{T}}]^{\mathcal{T}}$, where $\tilde{\mathbf{G}}_k \in \mathbb{F}_q^{L_k \times L}$, and $\tilde{\mathbf{g}}_{k,l_k}$ denotes the l_kth row of $\tilde{\mathbf{G}}_k$, $l_k = 1, 2, \ldots, L_k$. Generally, the recovery of various requested messages at each user requires the knowledge of $\tilde{\mathbf{G}}_{1\sim K}$. Because of the invertibility of $\mathbf{G}_{1\sim K}$, the messages $\mathbf{w}_{\gamma_m}$ requested by user m can be recovered by $\mathbf{w}_{\gamma_m} = \tilde{\mathbf{G}}_{\gamma_m} \otimes \mathbf{u}_{1\sim K}$, $\gamma_m \in \{1, 2, \ldots, M\}$. This paper focuses on the design of $\tilde{\mathbf{G}}_{1\sim K}$ (will be elaborated in Section IV-C). As such, the joint caching matrix $\mathbf{G}_{1\sim K}$ is straightforward.

Remark 1 (Related works on cache placement): In some caching networks, the messages requested by some users may be exclusive to a specific BS due to the privacy concern.[1] This implies that the same caching messages may not be allowed to share among different BSs. In this paper, we consider that the independent channel coding and modulation are performed on the caching messages of each BS. Therefore, differing from [16], [18], [19], the proposed cache placement does not require different BSs to share their caching messages. ■

In what follows, Sections IV and V consider the fronthaul between the central unit and BSs without rate constraint. Considering the rate constraint, Section VI investigates a caching scenario where some BSs cannot prefetch enough caching messages to fulfill the user requests.

B. Signal-Space Alignment (SSA) Enabled Content Delivery Phase

Each multi-antenna BS precodes their caching messages by a precoding matrix and broadcasts its precoded signals to all users simultaneously. We will not delve into the channel coding and modulation that are performed on each caching message vector at each BS (please refer to the nested lattice coding design and decoding algorithm in [26], [31] for details). As such, the N_R-dimensional received signal vector at user m is given by

$$\mathbf{y}_m = \sum_{k=1}^{K} \mathbf{H}_{k,m}\mathbf{P}_k\mathbf{u}_k + \mathbf{z}_m,\ m = 1, 2, \ldots, M, \tag{6}$$

where $\mathbf{P}_k$ denotes the precoding matrix at BS k with size of $N_T \times L_k$. Let $\mathbf{p}_{k,l_k}$ denote the l_kth column of $\mathbf{P}_k$. Then, the joint precoding matrix at all K BSs is $\mathbf{P}_{1\sim K} = [\mathbf{P}_1\ \mathbf{P}_2\ \ldots\ \mathbf{P}_K]$ with size of $N_T \times L$. Let $\mathbf{p}_l$ denote the lth column of $\mathbf{P}_{1\sim K}, l = 1, 2, \ldots, L$. Considering that the caching message vector at each BS has the unit power, we require that $\sum_{k=1}^{K} E\{\mathrm{tr}(\mathbf{P}_k^{\mathcal{T}}\mathbf{P}_k)\} \le P_T$.

Recall that user m requests $\mathbf{w}_{\gamma_m} \in \mathbb{F}_q^{L_{\gamma_m}}$ from the N_R-dimensional $\mathbf{y}_m$ that involves the superimposition of L caching messages from the BSs. In this paper, we consider that $L > N_R$ and $L_{\gamma_m} \le N_R$.[2] Define the set that collects the vectors corresponding to the *signal spaces* of the NC caching messages received at user m as $\mathcal{V}_m = \{\mathcal{V}_{1,m}, \mathcal{V}_{2,m}, \ldots, \mathcal{V}_{K,m}\}$, where the subset $\mathcal{V}_{k,m} = \{\mathbf{H}_{k,m}\mathbf{p}_{k,1}\ \mathbf{H}_{k,m}\mathbf{p}_{k,2} \ldots \mathbf{H}_{k,m}\mathbf{p}_{k,L_k}\}$ with the vector $\mathbf{H}_{k,m}\mathbf{p}_{k,l_k}$ corresponding to signal space of the NC caching message u_{k,l_k}, $k = 1, 2, \ldots, K$ and $l_k = 1, 2, \ldots, L_k$. Given $|\mathcal{V}_{k,m}| = L_k$, the size of signal space in $\mathcal{V}_{k,m}$ is L_k, where $|\mathcal{A}|$ denotes the cardinality of set $\mathcal{A}$. In total, $|\mathcal{V}_m|$ NC caching messages collide at each user, where $|\mathcal{V}_m| = |\mathcal{V}_{1,m}| + \ldots + |\mathcal{V}_{K,m}| = L$ denotes the total size of signal spaces. Note that any two vectors of $\mathcal{V}_m$ associated with two different signal spaces are linearly independent under the random fading channels almost surely.

Now it is clear that the size of signal spaces (i.e., L) exceeds the spatial dimension of the received signal at each user (i.e., N_R). Under the dimension constraint at each user, the precoders should be jointly designed to deliberately align signal spaces of desired NC caching messages at each user.

IV. Signal-Space Alignment for Linear NC Caching

In the content delivery phase, all BSs broadcast different NC caching message vectors to all users simultaneously, and the interference at each user comes from the transmission of requested messages by all the other users. Since each NC caching message is a linear combination of multiple user requests, it is not possible for each user to extract the requested messages from the caching message by interference alignment (IA) [27]. Following the SSA designs in [25], [28], we propose a new SSA pattern for the linear NC wireless caching to align the desired NC caching messages under various user requests. In general, the key idea of IA is to align the interference signals in the same dimension separable from the desired signals, such that the desired signals are distinguishable. The SSA for the wireless network coding advances IA by exploiting both structures of the desired signals and interference, corresponding to the NC operation. As Section IV-A will show, the desired caching messages are actually not the messages requested by the user but those involved in the inverse NC matrix.

Definition 1 (Signal-Space Alignment): Two different vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are said to be aligned in the same signal space if $\mathbf{v}_1 = c\mathbf{v}_2$ with a nonzero constant c. This paper uses $\mathbf{v}_1 \| \mathbf{v}_2$ to denote two aligned vectors $\mathbf{v}_1$ and $\mathbf{v}_2$. ■

Index Conversion: Suppose that a length-L column vector $\mathbf{v}_{1\sim K} = [\mathbf{v}_1^{\mathcal{T}}\ \mathbf{v}_2^{\mathcal{T}}\ \ldots\ \mathbf{v}_K^{\mathcal{T}}]^{\mathcal{T}}$. Let $\mathbf{v}_k$ be the kth length-L_k subvector of $\mathbf{v}_{1\sim K}$. Then, $L = L_1 + L_2 + \ldots + L_K$. For clarity, the l_kth entry of $\mathbf{v}_k$ corresponds to the $I(k, l_k)$th entry of $\mathbf{v}_{1\sim K}$, where

$$I(k, l_k) = \sum_{i=1}^{k-1} L_i + l_k, \tag{7}$$

for $k \in \{1, 2, \ldots, K\}$ and $l_k \in \{1, 2, \ldots, L_k\}$. Conversely, the tth entry of $\mathbf{v}_{1\sim K}$ corresponds to the $I_2^{\dagger}(t)$th entry of $\mathbf{v}_{I_1^{\dagger}(t)}$,

[1] In the content delivery network (CDN), each content owner such as a media company or an e-commerce vendor rents a group of servers owned by a CDN operator to deliver the requested contents to its end users. In this case, the servers rented by the content owner are exclusive to a specific group of users.

[2] For $L \le N_R$, each user can decode the L caching messages from the N_R-dimensional received signal vector. However, the caching scheme cannot achieve the maximum DoF. Meanwhile, if $L_{\gamma_m} > N_R$, the user cannot decode the length-L_{γ_m} requested messages within the N_R spatial dimension.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2020.3035607, IEEE Transactions on Communications

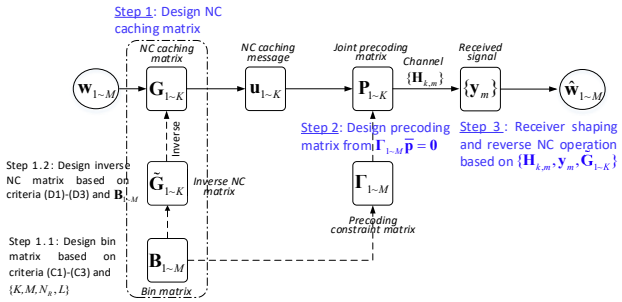


Fig. 3. The design procedure of matrices in Section IV.

as follows:

$$I_1^{\dagger}(t) = k, \text{ if } \sum_{i=1}^{k-1} L_i < t \leq \sum_{i=1}^{k} L_i, \tag{8}$$

$$I_2^{\dagger}(t) = t - \sum_{i=1}^{I_1^{\dagger}(t)-1} L_i. \tag{9}$$

The design procedure of matrices in the following subsections is illustrated in Fig. 3.

A. Bin Designs at Users

In this subsection, we design a bin matrix $\mathbf{B}_m$ over $\mathbb{F}_2$ to identify the desired NC caching messages that should be aligned at the user. Furthermore, $\mathbf{B}_m = [\mathbf{b}_{m,1}^{\mathcal{T}} \ \mathbf{b}_{m,2}^{\mathcal{T}} \ \dots \ \mathbf{b}_{m,N}^{\mathcal{T}}]^{\mathcal{T}}$ with $\mathbf{b}_{m,n} \in \mathbb{F}_2^L$ being the nth row vector of $\mathbf{B}_m$, $n = 1, 2, \dots, N$. We also determine the value of N, i.e., the number of bins at each user. Define $\mathbf{B}_{1\sim M} = [\mathbf{B}_1^{\mathcal{T}} \ \mathbf{B}_2^{\mathcal{T}} \dots \mathbf{B}_M^{\mathcal{T}}]^{\mathcal{T}}$ as the joint bin matrix. The SSA pattern for the linear NC caching network relies on the structure of $\mathbf{B}_{1\sim M}$.

Define a support set that collects the indices of nonzero entries in $\mathbf{b}_{m,n}$ as

$$\mathcal{J}(\mathbf{b}_{m,n}) = \big\{ j \in \{1, 2, \dots, L\} \big| b_{m,n}[j] = 1 \big\}, \tag{10}$$

where $b_{m,n}[j]$ denotes the jth entry of $\mathbf{b}_{m,n}$.

Suppose that N bins are created for each user to align the L messages within the N_R dimension. Generally, bin n at the user contains a set of aligned NC caching messages, as follows:

$$\mathcal{B}_{m,n} = \big\{ u_{k,l_k} | I(k, l_k) \in \mathcal{J}(\mathbf{b}_{m,n}), \exists k \in \{1, 2, \dots, K\}, \exists l_k \in \{1, 2, \dots, L_k\} \big\}, \tag{11}$$

for $n = 1, 2, \dots, N$ and $m = 1, 2, \dots, M$, where $I(k, l_k)$ is defined in (7). The aligned caching messages in each user's bin are so call the desired caching messages of this user. From (11), the caching messages in $\mathcal{B}_{m,n}$ correspond to the nonzero elements in $\mathbf{b}_{m,n}$, and $|\mathcal{B}_{m,n}| = |\mathcal{J}(\mathbf{b}_{m,n})|$. Further, $|\mathcal{B}_{m,n}| \leq L$, since $|\mathcal{J}(\mathbf{b}_{m,n})| \leq L$, $\forall m \in \{1, 2, \dots, M\}$ and $\forall n \in \{1, 2, \dots, N\}$. We remark that there exists a single vector occupying the unique signal space if $|\mathcal{B}_{m,n}| = 1$.

Due to the spatial dimension constraint at each user, the joint bin matrix $\mathbf{B}_{1\sim M}$ should be deliberately designed such that the user can decode its requested messages. Considering $MN_R \geq L$,[3] the design criteria of the joint bin matrix $\mathbf{B}_{1\sim M}$ for M users are as follows:

(C1) The Hamming weight[4] of every column in $\mathbf{B}_m$ is 1, $m = 1, 2, \dots, M$;

(C2) $N = \max\limits_{k \in \{1, \dots, K\}} L_k$;

(C3) $\mathbf{B}_{1\sim M}$ has the full rank of L in $\mathbb{F}_2$.

Let us interpret the design criteria in (C1)-(C3).

First, (C1) implies that any two distinct bins at each user cannot contain the same NC caching message. Thus, the signal space associated with each bin is unique. That is, $\mathcal{B}_{m,n} \cap \mathcal{B}_{m,n'} = \emptyset, \forall n \neq n', \forall n, n' \in \{1, 2, \dots, N\}$, where $\emptyset$ denotes the empty set. If (C1) fails, the NC caching messages in the two bins cannot be separated at the user, since they are aligned together.

Second, (C2) indicates that the number of bins at each user is $N = \max\limits_{\forall k \in \{1, \dots, K\}} L_k$. Given that $N \leq N_R$, all NC caching messages are aligned within the N-dimensional signal space. For the user m with the length of the requested message $L_{\gamma_m} = N$, it is straightforward that N bins are needed. For the user m' with $L_{\gamma_{m'}} < N$, N bins are still required since the user with $L_{\gamma_m} = N$ will interfere the user m', for $m' \neq m$. In addition, the maximum number of bins created at each user is N_R, i.e., $N \leq N_R$.

Third, (C3) guarantees that the column vectors of $\mathbf{B}_{1\sim M}$ are linearly independent over $\mathbb{F}_2$. Otherwise, it is possible that some column vectors in $\mathbf{B}_m$ can be expressed as the linear combinations of the remaining columns in $\mathbf{B}_m$ over $\mathbb{F}_2$, which contradicts with (C1). As Section IV-C further verifies, (C3) is required to design the full-rank inverse NC matrix over $\mathbb{F}_q$.

Furthermore, from (C1) and (C2), we obtain $\sum_{n=1}^{N} |\mathcal{B}_{m,n}| = L$, $m = 1, 2, \dots, M$. In other words, all NC caching messages are included in the bins of each user.

Theorem 1: Consider a $K \times M$ linear NC wireless caching network, where each BS and user are equipped with N_T and N_R antennas, respectively. Given that $MN_R \geq L$, $L > N_R$, and $L_{r_m} \leq N_R, \forall m$, there exists a joint bin matrix $\mathbf{B}_{1\sim M} \in \mathbb{F}_2^{MN \times L}$ satisfying (C1)-(C3).

The proof of Theorem 1 is given in Appendix. ■

Observation 1: The designed $\mathbf{B}_{1\sim M}$ in Theorem 1 is independent of γ_m for the worst-case caching scenario. In this case, the $\mathbf{B}_{1\sim M}$ keeps invariant as the user request varies. ■

Corollary 1: Following the statement of Theorem 1, there exists a joint bin matrix $\mathbf{B}_{1\sim M}$ satisfying (C1)-(C3) if and only if at most one of $L_{\gamma_1}, L_{\gamma_2}, \dots, L_{\gamma_M}$ is equal to N_R.

Proof: As (C2) implies, $N \leq N_R$. W.l.o.g., consider that $L_{\gamma_1} = N_R$. The sufficiency is proved by following Case 1 in Appendix. If $L_{\gamma_1} \geq N_R > L_{\gamma_2} \geq \dots \geq L_{\gamma_M}$, the design of $\mathbf{B}_{1\sim M}$ satisfying (C1)-(C3) indeed exists. The necessity follows Case 2 in Appendix, by using the proof of contradiction. That is, it is not possible to design the $\mathbf{B}_{1\sim M}$ satisfying (C1)-(C3) if we can find any other L_{γ_m} yielding

[3] If $MN_R < L$, it implies that there exists at least one user m that cannot decode its requested messages because of $L_{\gamma_m} > N_R$.

[4] The Hamming weight of a vector denotes the number of nonzero elements in this vector.

$L_{\gamma_m} = L_{\gamma_1}$ for $m \in \{2, 3, \ldots, M\}$. The proof is completed. ■

The corollary below shows that $\mathbf{B}_{1\sim M}$ has the block-wise full rank in $\mathbb{F}_2$.

Corollary 2: Consider a bin matrix $\mathbf{B}_{1\sim M}$ satisfying (C1)-(C3). Each submatrix $\mathbf{B}_m$ has the full rank of N in $\mathbb{F}_2$.

Proof: According to the interpretation of (C1), any distinct $\mathbf{b}_{m,n}$ and $\mathbf{b}_{m,n'}$ cannot have "1" in the same place, since $\mathcal{B}_{m,n} \cap \mathcal{B}_{m,n'} = \emptyset$, $n \neq n'$. Thus, it is not possible to find a nonzero vector $\alpha = [\alpha_1\ \alpha_2\ \ldots\ \alpha_N] \in \mathbb{F}_2^N$ such that $\alpha_1\mathbf{b}_{m,1} + \alpha_2\mathbf{b}_{m,2} + \ldots + \alpha_N\mathbf{b}_{m,N} = 0 \pmod 2$. Therefore, $\mathbf{B}_m$ has the full rank of N in $\mathbb{F}_2$. ■

In the following, we illustrate the bin matrix design with two toy examples.

Example 1 (Part I: bin matrix): Considers $K = M = 2$ and $N_T = N_R = 3$ [30]. The message vectors are $\mathbf{w}_1 = [w_{1,1}\ w_{1,2}\ w_{1,3}]^{\mathcal{T}}$ and $\mathbf{w}_2 = [w_{2,1}\ w_{2,2}]^{\mathcal{T}}$, where $\mathbf{w}_1 \in \mathbb{F}_q^3$ and $\mathbf{w}_2 \in \mathbb{F}_q^2$. Suppose that users 1 and 2 request $\mathbf{w}_1$ and $\mathbf{w}_2$ respectively, i.e., $\{\gamma_1, \gamma_2\} = \{1, 2\}$.

Cache placement phase: The specific designs of $\mathbf{G}_1$ and $\mathbf{G}_2$ will be given in Part III of Example 1. Then, BSs 1 and 2 store $\mathbf{u}_1 = \mathbf{G}_1 \otimes \mathbf{w}_{1\sim 2} = [u_{1,1}\ u_{1,2}\ u_{1,3}]^{\mathcal{T}}$ and $\mathbf{u}_2 = \mathbf{G}_2 \otimes \mathbf{w}_{1\sim 2} = [u_{2,1}\ u_{2,2}]^{\mathcal{T}}$ at their local caches, respectively.

Content delivery phase: From (6), the vector w.r.t. signal spaces of NC caching messages received at user m is $\mathcal{V}_m = \{\mathbf{H}_{1,m}\mathbf{p}_{1,1}\ \mathbf{H}_{1,m}\mathbf{p}_{1,2}\ \mathbf{H}_{1,m}\mathbf{p}_{1,3}\ \mathbf{H}_{2,m}\mathbf{p}_{2,1}\ \mathbf{H}_{2,m}\mathbf{p}_{2,2}\}$, $m = 1, 2$. It is seen that the size of signal spaces exceeds the spatial dimension of the received signal at each user, since $|\mathcal{V}_1| = |\mathcal{V}_2| = 5 > N_R = 3$. By Theorem 1, $\mathbf{B}_{1\sim 2}$ is designed as

$$\mathbf{B}_{1\sim 2} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{bmatrix} = \left[\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ \hline 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{array}\right]. \tag{12}$$

One can verify that $\mathbf{B}_{1\sim 2}$ in (12) satisfy (C1)-(C3). At user 1, 3 bins are created with respect to $\mathbf{B}_1$. First, from (10), $\mathcal{J}(\mathbf{b}_{1,1}) = \{1\}$, $\mathcal{J}(\mathbf{b}_{1,2}) = \{2, 4\}$, and $\mathcal{J}(\mathbf{b}_{1,3}) = \{3, 5\}$. Then, from (11), $\mathcal{B}_{1,1} = \{u_1\}$, $\mathcal{B}_{1,2} = \{u_2\ u_4\}$, and $\mathcal{B}_{1,3} = \{u_3\ u_5\}$. At user 2, 3 bins are created with respect to $\mathbf{B}_2$. First, $\mathcal{J}(\mathbf{b}_{2,1}) = \{1, 4\}$, $\mathcal{J}(\mathbf{b}_{2,2}) = \{2, 5\}$, and $\mathcal{J}(\mathbf{b}_{2,3}) = \{3\}$. Then, $\mathcal{B}_{2,1} = \{u_1\ u_4\}$, $\mathcal{B}_{2,2} = \{u_2\ u_5\}$, and $\mathcal{B}_{2,3} = \{u_3\}$. ■

Example 2 (Part I: Bin matrix): This example considers $K = M = 3$ and $N_T = 6, N_R = 3$. Consider $\mathbf{w}_1 = [w_{1,1}\ w_{1,2}\ w_{1,3}]^{\mathcal{T}}$, $\mathbf{w}_2 = [w_{2,1}\ w_{2,2}]^{\mathcal{T}}$, and $\mathbf{w}_3 = [w_{3,1}\ w_{3,2}]^{\mathcal{T}}$. Suppose that $\{\gamma_1, \gamma_2, \gamma_3\} = \{1, 2, 3\}$. By Theorem 1, $\mathbf{B}_{1\sim 3}$ is designed as

$$\mathbf{B}_{1\sim 3} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \\ \mathbf{B}_3 \end{bmatrix} = \left[\begin{array}{ccccccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ \hline 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ \hline 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{array}\right]. \tag{13}$$

B. SSA Enabled Precoding Design at BSs

In this subsection, we jointly design the precoding matrices of K BSs to perform the SSA characterized by the joint bin matrix. Suppose that user m requests $\mathbf{w}_{\gamma_m} \in \mathbb{F}_q^{L_{\gamma_m}}$. Let $\mathcal{J}(\mathbf{b}_{\gamma_m,n}) = \{j_1, j_2, \ldots, j_{|\mathcal{B}_{\gamma_m,n}|}\}$. To align the NC caching messages in $\mathcal{B}_{\gamma_m,n}$ of (11), $\mathbf{P}_{1\sim K}$ must satisfy

$$\mathbf{H}_{I_1^\dagger(j_1),m}\mathbf{p}_{j_1} \| \mathbf{H}_{I_1^\dagger(j_2),m}\mathbf{p}_{j_2} \ldots \| \mathbf{H}_{I_1^\dagger(j_{|\mathcal{B}_{\gamma_m,n}|}),m}\mathbf{p}_{j_{|\mathcal{B}_{\gamma_m,n}|}} \tag{14}$$

From (8) and (9), we can alternatively express (14) as

$$\begin{aligned} &\mathbf{H}_{I_1^\dagger(j_1),m}\mathbf{p}_{I_1^\dagger(j_1),I_2^\dagger(j_1)} \| \mathbf{H}_{I_1^\dagger(j_2),m}\mathbf{p}_{I_1^\dagger(j_2),I_2^\dagger(j_2)} \\ &\ldots \| \mathbf{H}_{I_1^\dagger(j_{|\mathcal{B}_{\gamma_m,n}|}),m}\mathbf{p}_{I_1^\dagger(j_{|\mathcal{B}_{\gamma_m,n}|}),I_2^\dagger(j_{|\mathcal{B}_{\gamma_m,n}|})}. \end{aligned} \tag{15}$$

From Definition 1, (14) indicates that

$$\begin{aligned} c_1\mathbf{H}_{I_1^\dagger(j_1),m}\mathbf{p}_{j_1} &= c_2\mathbf{H}_{I_1^\dagger(j_2),m}\mathbf{p}_{j_2} = \ldots \\ &= c_{|\mathcal{B}_{m,n}|}\mathbf{H}_{I_1^\dagger(j_{|\mathcal{B}_{\gamma_m,n}|}),m}\mathbf{p}_{j_{|\mathcal{B}_{\gamma_m,n}|}}, \end{aligned} \tag{16}$$

for some constants $c_1, c_2, \ldots, c_{|\mathcal{B}_{\gamma_m,n}|}$. For the binary NC caching matrix in this paper, $c_1 = c_2 = \ldots = c_{|\mathcal{B}_{\gamma_m,n}|} = 1$.

Considering all bins at all users, we can rewrite (16) in a matrix form:

$$\mathbf{\Gamma}_{1\sim M}\bar{\mathbf{p}} = \mathbf{0}, \tag{17}$$

where $\mathbf{\Gamma}_{1\sim M}$ is referred to as the *joint precoding constraint matrix*, and $\bar{\mathbf{p}} = [\mathbf{p}_{1,1}^{\mathcal{T}}, \mathbf{p}_{1,2}^{\mathcal{T}}, \ldots, \mathbf{p}_{1,L_k}^{\mathcal{T}}, \ldots, \mathbf{p}_{K,1}^{\mathcal{T}}, \mathbf{p}_{K,2}^{\mathcal{T}}, \ldots, \mathbf{p}_{K,L_K}^{\mathcal{T}}]^{\mathcal{T}}$ denotes the vectorized version of $\mathbf{P}_{1\sim K}$ with $\mathbf{p}_{k,l_k}$ defined in (6). The length of $\bar{\mathbf{p}}$ is $(L_1 + L_2 + \ldots + L_K)N_T = LN_T$. The intuition of introducing $\mathbf{\Gamma}_{1\sim M}$ is to formulate a constraint under which the precoding matrices of BSs exist.

Let us clarify the structure of $\mathbf{\Gamma}_{1\sim M}$ based on (16). First, we can decompose $\mathbf{\Gamma}_{1\sim M}$ into

$$\mathbf{\Gamma}_{1\sim M} = [\mathbf{\Gamma}_1^{\mathcal{T}}\ \mathbf{\Gamma}_2^{\mathcal{T}} \ldots \mathbf{\Gamma}_M^{\mathcal{T}}]^{\mathcal{T}}, \tag{18}$$

where the mth row submatrix $\mathbf{\Gamma}_m$ is the precoding constraint matrix for the user m according to $\mathbf{B}_m$. Plugging (18) into (17), the precoding constraint matrix for user m yields

$$\mathbf{\Gamma}_m\bar{\mathbf{p}} = \mathbf{0}. \tag{19}$$

Note that the number of bins created at each user is N by (C2). Then, $\mathbf{\Gamma}_m$ is expressed as

$$\mathbf{\Gamma}_m = [\mathbf{\Gamma}_{m,1}^{\mathcal{T}}\ \mathbf{\Gamma}_{m,2}^{\mathcal{T}}\ \ldots\ \mathbf{\Gamma}_{m,N}^{\mathcal{T}}]^{\mathcal{T}}, \tag{20}$$

where $\mathbf{\Gamma}_{m,n}$ is the precoding constraint matrix induced by the bin n of user m, $n = 1, 2, \ldots, N$. More specific, we have

$$\mathbf{\Gamma}_{m,n} = \begin{bmatrix} \cdots & \mathbf{H}_{I_1^\dagger(j_1),m} & \cdots & \cdots & \mathbf{H}_{I_1^\dagger(j_2),m} & \cdots \\ \cdots & \mathbf{H}_{I_1^\dagger(j_1),m} & \cdots & \cdots & \cdots & \mathbf{H}_{I_1^\dagger(j_3),m} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdots & \mathbf{H}_{I_1^\dagger(j_1),m} & \cdots & \mathbf{H}_{I_1^\dagger(j_{|\mathcal{B}_{\gamma_m,n}|}),m} & \cdots & \cdots \end{bmatrix}, \tag{21}$$

where $\mathcal{J}(\mathbf{b}_{\gamma_m,n}) = \{j_1, j_2, \ldots, j_{|\mathcal{B}_{\gamma_m,n}|}\}$ is defined in (10). It is emphasized that $\mathbf{\Gamma}_{m,n}$ is a sparse matrix with two nonzero submatrices only in each row block and zero submatrices elsewhere. Therefore, each $\mathbf{\Gamma}_{m,n}$ has the size of

$(|\mathcal{B}_{\gamma_m,n}|-1)N_R \times LN_T$. Taking all N bins of user m into consideration, $\mathbf{\Gamma}_m$ has the size of $(L-N)N_R \times LN_T$, since $|\mathcal{B}_{\gamma_m}| = \sum_{n=1}^{N} |\mathcal{B}_{\gamma_m,n}| = L$ by (C2).

It is possible that some bin n' contains a single NC caching message. In this case, from (14) and (21), the corresponding $\mathbf{\Gamma}_{m,n'}$ contains a single nonzero submatrix only. However, this $\mathbf{\Gamma}_{m,n'}$ does not impose any constraint for the SSA of user m, and hence can be safely removed from $\mathbf{\Gamma}_m$. Then, $\mathbf{\Gamma}_m$ is obtained by collecting the remaining $\mathbf{\Gamma}_{m,n}$ in $\mathbf{\Gamma}_m$, $\forall n \neq n', n \in \{1,2,\ldots,N\}$.

Theorem 2: Given a joint bin matrix $\mathbf{B}_{1\sim M}$ satisfying (C1)-(C3), there exists at least one solution of $\mathbf{P}_{1\sim K}$ if

$$LN_T > (L-N)MN_R. \tag{22}$$

Proof: From (19), $\bar{\mathbf{p}}$ exists as long as the null space of $\mathbf{\Gamma}_{1\sim M}$ is not empty. It is straightforward from (21) that $\mathbf{\Gamma}_{1\sim M}$ has the size of $(L-N)MN_R \times LN_T$. Suppose that the rank of $\mathbf{\Gamma}_{1\sim M}$ is $\mathrm{rank}(\mathbf{\Gamma}_{1\sim M})$. According to the rank-nullity theorem [29], the nullity of $\mathbf{\Gamma}_{1\sim M}$ is $\mathrm{nul}(\mathbf{\Gamma}_{1\sim M}) = LN_T - \mathrm{rank}(\mathbf{\Gamma}_{1\sim M})$. If (22) is satisfied, there always exists at least one solution of $\bar{\mathbf{p}}$, since $\mathrm{rank}(\mathbf{\Gamma}_{1\sim M}) \leq (L-N)MN_R$ yielding $\mathrm{nul}(\mathbf{\Gamma}_{1\sim M}) > 0$. The proof is completed. ■

Observation 2: For any request γ_m, $\mathbf{\Gamma}_m$ is induced by the bin $\mathbf{B}_{\gamma_m}$. Therefore, different user requests lead to different joint precoding matrices according to (21).

Example 1 (Part II: precoding constraint matrix) Following the bin matrix (12) in Part I of Example 1, the joint precoding constraint matrix under $\{\gamma_1,\gamma_2\} = \{1,2\}$ is given by

$$\mathbf{\Gamma}_{1\sim 2} = \begin{bmatrix} \mathbf{\Gamma}_1 \\ \mathbf{\Gamma}_2 \end{bmatrix} = \left[\begin{array}{ccccc} 0 & \mathbf{H}_{1,1} & 0 & \mathbf{H}_{2,1} & 0 \\ 0 & 0 & \mathbf{H}_{1,1} & 0 & \mathbf{H}_{2,1} \\ \hline \mathbf{H}_{1,2} & 0 & 0 & \mathbf{H}_{2,2} & 0 \\ 0 & \mathbf{H}_{1,2} & 0 & 0 & \mathbf{H}_{2,2} \end{array}\right] \tag{23}$$

and $\bar{\mathbf{p}} = [\mathbf{p}_{1,1}^{\mathcal{T}}, \mathbf{p}_{1,2}^{\mathcal{T}}, \mathbf{p}_{1,3}^{\mathcal{T}}, \mathbf{p}_{2,1}^{\mathcal{T}}, \mathbf{p}_{2,2}^{\mathcal{T}}]^{\mathcal{T}}$. It can be verified that the SSA at users 1 and 2 yields $\mathbf{\Gamma}\bar{\mathbf{p}} = \mathbf{0}$. We remark that $\mathbf{\Gamma}_{1\sim 2}$ in (23) does not impose any constraint on the bins $\mathbf{b}_{1,1}$ and $\mathbf{b}_{2,3}$ each with a single message.

Correspondingly, the SSA in bins 2 and 3 of user 1 yields $\mathbf{H}_{1,1}\mathbf{p}_2 \| \mathbf{H}_{2,1}\mathbf{p}_4$ (i.e., $\mathbf{H}_{1,1}\mathbf{p}_{1,2} \| \mathbf{H}_{2,1}\mathbf{p}_{2,1}$) and $\mathbf{H}_{1,1}\mathbf{p}_3 \| \mathbf{H}_{2,1}\mathbf{p}_5$ (i.e., $\mathbf{H}_{1,1}\mathbf{p}_{1,3} \| \mathbf{H}_{2,1}\mathbf{p}_{2,2}$), respectively, while bin 1 of user 1 contains a single signal vector of $\mathbf{H}_{1,1}\mathbf{p}_1$ (i.e., $\mathbf{H}_{1,1}\mathbf{p}_{1,1}$). Likewise, the SSA in bins 1 and 2 of user 2 is $\mathbf{H}_{1,2}\mathbf{p}_1 \| \mathbf{H}_{2,2}\mathbf{p}_4$ (i.e., $\mathbf{H}_{1,2}\mathbf{p}_{1,1} \| \mathbf{H}_{2,2}\mathbf{p}_{2,1}$) and $\mathbf{H}_{1,2}\mathbf{p}_2 \| \mathbf{H}_{2,2}\mathbf{p}_5$ (i.e., $\mathbf{H}_{1,2}\mathbf{p}_{1,2} \| \mathbf{H}_{2,2}\mathbf{p}_{2,2}$), respectively, while bin 3 of user 2 contains a single signal vector of $\mathbf{H}_{1,2}\mathbf{p}_3$ (i.e., $\mathbf{H}_{1,2}\mathbf{p}_{1,3}$). By Theorem 2, we can find at least one precoding matrix, since the size of $\mathbf{\Gamma}_{1\sim 2}$ is 12×15.

Consider $\{\gamma_1,\gamma_2\} = \{2,1\}$. From (21), the joint precoding constraint matrix becomes

$$\mathbf{\Gamma}_{1\sim 2} = \begin{bmatrix} \mathbf{\Gamma}_1 \\ \mathbf{\Gamma}_2 \end{bmatrix} = \left[\begin{array}{ccccc} \mathbf{H}_{1,1} & 0 & 0 & \mathbf{H}_{2,1} & 0 \\ 0 & \mathbf{H}_{1,1} & 0 & 0 & \mathbf{H}_{2,1} \\ \hline 0 & \mathbf{H}_{1,2} & 0 & 0 & 0 \\ 0 & 0 & \mathbf{H}_{1,2} & 0 & \mathbf{H}_{2,2} \end{array}\right]. \tag{24}$$

The joint precoding matrix also exists under this request. ■

Example 2 (Part II: precoding constraint matrix) Following the bin matrix (13) in Part I of Example 2, the joint precoding constraint matrix under $\{\gamma_1,\gamma_2,\gamma_3\} = \{1,2,3\}$ is given by

$$\mathbf{\Gamma}_{1\sim 3} = \begin{bmatrix} \mathbf{\Gamma}_1 \\ \mathbf{\Gamma}_2 \\ \mathbf{\Gamma}_3 \end{bmatrix} = \left[\begin{array}{ccccccc} 0 & \mathbf{H}_{1,1} & 0 & \mathbf{H}_{2,1} & 0 & 0 & 0 \\ 0 & \mathbf{H}_{1,1} & 0 & 0 & 0 & \mathbf{H}_{3,1} & 0 \\ 0 & 0 & \mathbf{H}_{1,1} & 0 & \mathbf{H}_{2,1} & 0 & 0 \\ 0 & 0 & \mathbf{H}_{1,1} & 0 & 0 & 0 & \mathbf{H}_{3,1} \\ \hline \mathbf{H}_{1,2} & 0 & 0 & H_{2,2} & 0 & 0 & 0 \\ 0 & \mathbf{H}_{1,2} & 0 & 0 & \mathbf{H}_{2,2} & 0 & 0 \\ 0 & \mathbf{H}_{1,2} & 0 & 0 & 0 & \mathbf{H}_{3,2} & 0 \\ 0 & 0 & \mathbf{H}_{1,2} & 0 & 0 & 0 & \mathbf{H}_{3,2} \\ \hline \mathbf{H}_{1,3} & 0 & 0 & \mathbf{H}_{2,3} & 0 & 0 & 0 \\ \mathbf{H}_{1,3} & 0 & 0 & 0 & 0 & \mathbf{H}_{3,3} & 0 \\ 0 & \mathbf{H}_{1,3} & 0 & 0 & \mathbf{H}_{2,3} & 0 & 0 \\ 0 & \mathbf{H}_{1,3} & 0 & 0 & 0 & 0 & \mathbf{H}_{3,3} \end{array}\right] \tag{25}$$

and $\bar{\mathbf{p}} = [\mathbf{p}_{1,1}^{\mathcal{T}}, \mathbf{p}_{1,2}^{\mathcal{T}}, \mathbf{p}_{1,3}^{\mathcal{T}}, \mathbf{p}_{2,1}^{\mathcal{T}}, \mathbf{p}_{2,2}^{\mathcal{T}}, \mathbf{p}_{3,1}^{\mathcal{T}}, \mathbf{p}_{3,2}^{\mathcal{T}}]^{\mathcal{T}}$. By Theorem 2, the precoding matrix exists, since the size of $\mathbf{\Gamma}_{1\sim 3}$ is 36×42.

C. Joint NC Caching Matrix Design

Let us detail the design of the joint NC caching matrix $\mathbf{G}_{1\sim K}$. Given $\mathbf{B}_{1\sim M} \in \mathbb{F}_2^{MN\times L}$ and $M = K$, the design criteria of the inverse NC matrix $\tilde{\mathbf{G}}_{1\sim K} \in \mathbb{F}_2^{L\times L}$ are given as follows:

(D1) For BS k with $L_k = N$, $\tilde{\mathbf{G}}_k = \mathbf{B}_k$.

(D2) For BS k with $L_k < N$, $\tilde{\mathbf{G}}_k$ is a submatrix of $\mathbf{B}_k$ by selecting L_k rows of $\mathbf{B}_k$.

(D3) $\tilde{\mathbf{G}}_{1\sim K}$ has the full rank of L in $\mathbb{F}_2$.

For any $\mathbf{B}_{1\sim M}$ satisfying (C1)-(C3), we can verify that there exists at least one $\tilde{\mathbf{G}}_{1\sim K}$ satisfying (D1)-(D3), as follows:

For the case stated in (D1), the design of $\tilde{\mathbf{G}}_k$ is straightforward. For the case stated in (D2), suppose that the row indices of $\mathbf{B}_k$ taken by $\tilde{\mathbf{G}}_k$ are denoted by $\{l_1, l_2, \ldots, l_{L_k}\}$. Notably, the rows cannot be arbitrarily chosen from $\mathbf{B}_k$ to construct $\tilde{\mathbf{G}}_k$ in (D2). This is because the selected rows in $\mathbf{B}_k$ may be linearly dependent on $\mathbf{B}_{k'}$ over $\mathbb{F}_2$ for $k \neq k'$, resulting in a rank deficient $\tilde{\mathbf{G}}_{1\sim K}$ over $\mathbb{F}_2$. Thus, (D3) is further needed to guarantee that $\tilde{\mathbf{G}}_{1\sim K}$ has full rank over $\mathbb{F}_2$. Then, from (C3), the full rank $\tilde{\mathbf{G}}_k$ over $\mathbb{F}_2$ indeed exists, since $\tilde{\mathbf{G}}_{1\sim K} \subseteq \mathbf{B}_{1\sim K}$.

By Corollary 2, we can deduce that $\tilde{\mathbf{G}}_{1\sim K}$ also has the block-wise full rank property. That is, each submatrix $\tilde{\mathbf{G}}_k$ has full rank of L_k over $\mathbb{F}_q$. In addition, the Hamming weight of every column in $\tilde{\mathbf{G}}_k$ is also 1, since $\tilde{\mathbf{G}}_k \subseteq \mathbf{B}_k$.

For any $\tilde{\mathbf{G}}_{1\sim K}$ satisfying (D1)-(D3), we can identify its inverse matrix over $\mathbb{F}_q$, i.e., the joint NC caching matrix $\mathbf{G}_{1\sim K}$. Thus, the bin design is the core of the linear NC wireless caching, which bridges the cache placement and content delivery via the linear wireless network coding.

Observation 3: In the worst-case caching, the design of $\tilde{\mathbf{G}}_{1\sim K}$ in (D1)-(D3) for any user request can also be used under the other user requests. As such, the joint NC caching matrix induced by $\tilde{\mathbf{G}}_{1\sim K}$ keeps unchanged for distinct user requests, consistent with Observation 1.

From Observations 2 and 3, the design of NC caching functions in the placement phase is unaware of the future

user requests, but the precoding design in the delivery phase depends on the user requests.

Example 1 (Part III: joint NC caching matrix) Let us revisit $\mathbf{B}_{1\sim 2}$ in (12). From (D1), $\tilde{\mathbf{G}}_1 = \mathbf{B}_1$ for BS 1 with $L_1 = N = 3$. From (D2), for BS 2 with $L_2 = 2$, we select the first 2 rows of $\mathbf{B}_2$ to form $\tilde{\mathbf{G}}_2$. Then, the inverse NC matrix is given by

$$\tilde{\mathbf{G}}_{1\sim 2} = \begin{bmatrix} \tilde{\mathbf{G}}_1 \\ \tilde{\mathbf{G}}_2 \end{bmatrix} = \left[\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ \hline 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{array}\right]. \tag{26}$$

It can be verified that $\tilde{\mathbf{G}}_{1\sim 2}$ has the full rank of 5 in $\mathbb{F}_q$.

Then, the joint NC caching matrix $\mathbf{G}$, being the inverse matrix of $\tilde{\mathbf{G}}$ over $\mathbb{F}_q$, is given by

$$\mathbf{G}_{1\sim 2} = \begin{bmatrix} \mathbf{G}_1 \\ \mathbf{G}_2 \end{bmatrix} = \left[\begin{array}{ccccc} 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & q-1 & 0 \\ 1 & 1 & 1 & q-1 & q-1 \\ \hline q-1 & 0 & 0 & 1 & 0 \\ q-1 & q-1 & 0 & 1 & 1 \end{array}\right]. \tag{27}$$

In the cache placement phase, BSs 1 and 2 follow (3) to prefecth $\mathbf{u}_1 = \mathbf{G}_1 \otimes \mathbf{w}_{1\sim 2} = [w_{1,1}, w_{1,1} \oplus w_{1,2} \oplus (q-1) \otimes w_{2,1}, w_{1,1} \oplus w_{1,2} \oplus w_{1,3} \oplus (q-1) \otimes w_{2,1} \oplus (q-1) \otimes w_{2,2}]^{\mathcal{T}}$ and $\mathbf{u}_2 = \mathbf{G}_2 \otimes \mathbf{w}_{1\sim 2} = [(q-1) \otimes w_{1,1} \oplus w_{2,1}, (q-1) \otimes w_{1,1} \oplus (q-1) \otimes w_{1,2} \oplus w_{2,1} \oplus w_{2,2}]^{\mathcal{T}}$ at their local caches, respectively. ■

Example 2 (Part III: joint NC caching matrix) Based on $\mathbf{B}_{1\sim 2}$ in (13), we have

$$\tilde{\mathbf{G}}_{1\sim 3} = \begin{bmatrix} \tilde{\mathbf{G}}_1 \\ \tilde{\mathbf{G}}_2 \\ \tilde{\mathbf{G}}_3 \end{bmatrix} = \left[\begin{array}{ccccccc} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ \hline 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ \hline 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 \end{array}\right]. \tag{28}$$

Then, $\mathbf{G}_{1\sim 3}$ can also be determined as the inverse matrix of $\tilde{\mathbf{G}}_{1\sim 3}$ over $\mathbb{F}_q$. ■

D. Receiver Shaping and Reverse NC Operation

By the SSA in Section IV-B, the desired NC caching messages are aligned at the bins of each user. Then, the zero-forcing receiver shaping is used to decouple the bins of each user, assuming that $\tilde{\mathbf{G}}_{1\sim K}$ is available at all users[5].

To decouple the bins, $\mathbf{y}_m$ of user m goes through a zero-forcing filter $\mathbf{Q}_m$ as below:

$$\mathbf{r}_m = \left[r_{m,1}\ r_{m,2} \dots\ r_{m,L_{\gamma_m}}\right]^{\mathcal{T}} = \mathbf{Q}_m \mathbf{y}_m, \tag{29}$$

where $\mathbf{Q}_m = [\mathbf{q}_{m,1}^{\mathcal{T}}\ \mathbf{q}_{m,2}^{\mathcal{T}}\ \dots\ \mathbf{q}_{m,L_{\gamma_m}}^{\mathcal{T}}]^{\mathcal{T}}$ denotes the shaping matrix with $\mathbf{q}_{m,l_m}$ being the l_mth row of $\mathbf{Q}_m$, $l_m = 1, 2, \dots, L_{\gamma_m}$. By the filter $\mathbf{q}_{m,l_m}$, $\mathbf{y}_m$ is projected onto the subspace perpendicular to that spanned by all other bins except bin l_m at user m. Therefore, we have

$$\mathbf{q}_{m,l_m} \mathbf{H}_{I_1^{\dagger}(j),m} \mathbf{p}_j = 0,\ \forall j \in \mathcal{J}(\mathbf{b}_{\gamma_m,l'_m}), \\ l'_m \in \{1, 2, \dots, L_{\gamma_m}\}, l_m \neq l'_m. \tag{30}$$

Note that each bin occupies a unique signal space according to the SSA pattern in (C1) of Section IV-A. Thus, r_{m,l_m} of user m corresponds to the signal space in the bin l_m only, and the interference from all other bins is eliminated. To null out the interference, it is sufficient for $\mathbf{q}_{m,l_m}$ to be orthogonal to any single vector in the bin l'_m, $l_m \neq l'_m$, since all vectors in the same bin are aligned by the SSA in (14). After the decoupling of bins, user m carries out the reverse NC operation to decode its requested $\mathbf{w}_{\gamma_m}$.

Consider that user m attempts to decode w_{γ_m,l_m}. From (29) and (30), $r_{m,l_m} = \mathbf{q}_{m,l_m}\mathbf{y}_m = \eta_{m,l_m} \sum_{\forall l \in \mathcal{J}(\mathbf{b}_{\gamma_m,l_m})} u_{\gamma_m,l}$,[6] where $\eta_{m,l_m} = \mathbf{q}_{m,l_m}\mathbf{H}_{I_1^{\dagger}(j),m}\mathbf{p}_j$ with $j \in \mathcal{J}(\mathbf{b}_{\gamma_m,l_m})$ according to the SSA in (16). That is, after the filtering, only the caching messages in bin $\mathbf{b}_{\gamma_m,l_m}$ are retained. Furthermore, as Section IV-C shows, $\mathbf{b}_{\gamma_m,l_m}$ is selected to be the l_mth row of $\tilde{\mathbf{G}}_{\gamma_m}$, i.e., $\tilde{\mathbf{g}}_{\gamma_m,l_m} = \mathbf{b}_{\gamma_m,l_m}$. Therefore, $r_{m,l_m} = \sum_{\forall l \in \mathcal{J}(\tilde{\mathbf{g}}_{\gamma_m,l_m})} u_{\gamma_m,l}$. Given this superimposed signal r_{m,l_m}, user m carries out the reverse NC operation to decode $w_{\gamma_m,l_m} = \eta_{m,l_m} \oplus_{\forall l \in \mathcal{J}(\tilde{\mathbf{g}}_{\gamma_m,l_m})} u_{\gamma_m,l}$, where the equality holds since $w_{\gamma_m,l_m} = \tilde{\mathbf{g}}_{\gamma_m,l_m} \otimes \mathbf{u}_{1\sim M} = \oplus_{\forall l \in \mathcal{J}(\tilde{\mathbf{g}}_{\gamma_m,l_m})} u_{\gamma_m,l}$. We remark that the reverse NC operation theoretically follows the compute-and-forward system [26]. In practice, the reverse NC operation can be realized by the modulation-coded physical-layer network coding scheme with irregular repeat-accumulate coding and iterative believe propagation decoding [31]. Let $\hat{\mathbf{w}}_{\gamma_m}$ be the estimated version of $\mathbf{w}_{\gamma_m}$. As (2) indicates, a decoding error occurs if $[\mathbf{w}_{\gamma_1}, \mathbf{w}_{\gamma_2}, \dots, \mathbf{w}_{\gamma_M}] \neq [\hat{\mathbf{w}}_{\gamma_1}, \hat{\mathbf{w}}_{\gamma_2}, \dots, \hat{\mathbf{w}}_{\gamma_M}]$. The decoding error of each user vanishes as the length of coding block at each BS goes to large.

Example 1 (Part IV: receiver shaping) The precoding constraint matrix and the inverse NC matrix follow (23) and (26) respectively. To decode $w_{1,1}$, $\mathbf{y}_1$ is projected to the subspace, perpendicular to that span by $\{\mathbf{H}_{1,1}\mathbf{p}_{1,2}, \mathbf{H}_{1,1}\mathbf{p}_{1,3}\}$. To decode $w_{1,2}$, $\mathbf{y}_1$ is projected to the subspace, perpendicular to that span by $\{\mathbf{H}_{1,1}\mathbf{p}_{1,1}, \mathbf{H}_{1,1}\mathbf{p}_{1,3}\}$. To decode $w_{1,3}$, $\mathbf{y}_1$ is projected to the subspace, perpendicular to that span by $\{\mathbf{H}_{1,1}\mathbf{p}_{1,1}, \mathbf{H}_{1,1}\mathbf{p}_{1,2}\}$. To decode $w_{2,1}$, $\mathbf{y}_2$ is projected to the subspace, perpendicular to that span by $\{\mathbf{H}_{1,2}\mathbf{p}_{1,2}, \mathbf{H}_{1,2}\mathbf{p}_{1,3}\}$. To decode $w_{2,2}$, $\mathbf{y}_2$ is projected to the subspace, perpendicular to that span by $\{\mathbf{H}_{1,2}\mathbf{p}_{1,1}, \mathbf{H}_{1,2}\mathbf{p}_{1,3}\}$. ■

Example 1 (Part V: reverse NC operation) Consider that user 1 attempts to recover $w_{1,2}$. First, the noise-free received signal at user 1 is $\mathbf{y}_1 = \mathbf{H}_{1,1}\mathbf{p}_{1,1}u_{1,1} + \mathbf{H}_{1,1}\mathbf{p}_{1,2}u_{1,2} + \mathbf{H}_{1,1}\mathbf{p}_{1,3}u_{1,3} + \mathbf{H}_{2,1}\mathbf{p}_{2,1}u_{2,1} + \mathbf{H}_{2,1}\mathbf{p}_{2,2}u_{2,2}$. After the receiver shaping in Part IV of Example 1, $r_{1,2} = \mathbf{q}_{1,2}\mathbf{y}_1 = \mathbf{q}_{1,2}\mathbf{H}_{1,1}\mathbf{p}_{1,2}u_{1,2} + \mathbf{q}_{1,2}\mathbf{H}_{2,1}\mathbf{p}_{2,1}u_{2,1}$. Furthermore, $r_{1,2}/\eta_{1,2} = u_{1,2} + u_{2,1}$, where $\eta_{1,2} = \mathbf{q}_{1,2}\mathbf{H}_{1,1}\mathbf{p}_{1,2} = \mathbf{q}_{1,2}\mathbf{H}_{2,1}\mathbf{p}_{2,1}$. Based on this superimposed signal, user 1 decodes $w_{1,2} = u_{1,2} \oplus u_{2,1} = \big(w_{1,1} \oplus w_{1,2} \oplus (q-1) \otimes w_{2,1}\big) \oplus$

[5]To retrieve the messages under different requests, it is essential for the requesting user to know $\tilde{\mathbf{G}}_{1\sim K}$. Therefore, we assume that $\tilde{\mathbf{G}}_{1\sim K}$ can be reliably transmitted to users. Since the pre-assigned $\mathbf{G}_{1\sim K}$ is compatible with different user requests under different channel realizations, the transmission overhead induced by $\tilde{\mathbf{G}}_{1\sim K}$ is negligible.

[6]For ease of interpretation, we consider a noise-free reception.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2020.3035607, IEEE Transactions on Communications

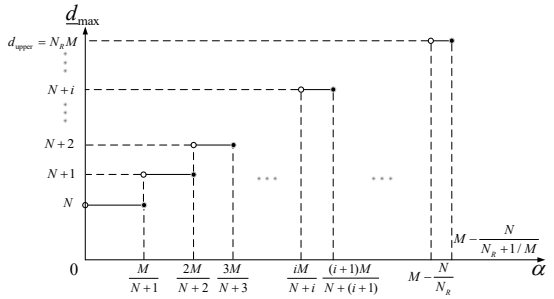


Fig. 4. Achievable $\underline{d}_{\max}$ of the proposed $K \times M$ wireless caching scheme under $\alpha \geq M/K$.

$\big((q-1) \otimes w_{1,1} \oplus w_{2,1}\big)$. The decoding for other messages in users 1 and 2 follows similarly. ■

Therefore, Parts I-V of Example 1 show that the requests of users 1 and 2 can be fulfilled by the proposed caching scheme in the worst-case caching scenario.

V. Sum DoF Analysis

Theorem 3: Consider the worst-case caching scenario. The proposed wireless caching network can achieve the maximum sum DoF as

$$\underline{d}_{\max} = \max_{L<\Delta, L\in\mathbb{Z}^+} L, \tag{31}$$

where $\Delta = \frac{MN_RN}{MN_R-N_T}$ and $\mathbb{Z}^+$ denotes the set of positive integers.

Proof: Given a $\tilde{\mathbf{G}}$ satisfying (D1)-(D3), by Theorem 2 the joint precoding matrix $\mathbf{P}_{1\sim K}$ exists if (22) is satisfied. Then, the SSA is guaranteed if $L < \Delta = \frac{MNN_R}{MN_R-N_T}, \exists L \in \mathbb{Z}^+$. Thus, the proposed scheme can achieve the sum DoF of $\underline{d} = L$. To maximize the $\underline{d}$, L should be as large as possible, as long as $L < \Delta$. Therefore, $\underline{d}$ reaches to the maxima if (31) is satisfied. ■

Corollary 3: Following the statement of Theorem 3, the maximum sum DoF $\underline{d}_{\max}$ is

$$\underline{d}_{\max} = \lceil \frac{MN_RN}{MN_R-N_T} \rceil - 1 = N + \lceil \frac{N_TN}{MN_R-N_T} \rceil - 1, \tag{32}$$

where $\lceil x \rceil$ denotes the ceiling function that maps x to the smallest integer larger than or equal to x.

Proof: By Theorem 3, if $\frac{MNN_R}{MN_R-N_T}$ is an integer, then $\underline{d}_{\max} = \frac{MNN_R}{MN_R-N_T} - 1$. If $\frac{MNN_R}{MN_R-N_T}$ is fractional and not an integer, then $\underline{d}_{\max} = \lfloor \frac{MNN_R}{MN_R-N_T} \rfloor$. To sum up, $\underline{d}_{\max} = \lceil \frac{MNN_R}{MN_R-N_T} \rceil - 1$. ■

Corollary 4: The maximum sum DoF $\underline{d}_{\max}$ in (31) changes as a piecewise step fashion as $\alpha = N_T/N_R$ varies: $\underline{d}_{\max} = N + i - 1,\ \alpha \in \left(\frac{(i-1)M}{N+(i-1)}, \frac{iM}{N+i} \right]$, for $i = \mathbb{Z}^+$.

Proof: Given that $\alpha = N_T/N_R$, we rewrite (31) as $\underline{d}_{\max} = N-1+\lceil \frac{\alpha N}{M-\alpha} \rceil$. It can be verified that $\lceil \frac{\alpha N}{M-\alpha} \rceil = i-1$ at $\alpha = \frac{(i-1)M}{N+i-1}$, $i \in \mathbb{Z}^+$. As α becomes slightly larger than $\frac{(i-1)M}{N+i-1}$, $\lceil \frac{\alpha N}{M-\alpha} \rceil = i$ and $\underline{d}_{\max}$ rises to $N+i-1$, as long as $\alpha \in \left(\frac{(i-1)M}{N+(i-1)}, \frac{iM}{N+i} \right]$. ■

To illustrate Corollary 4, Fig. 4 shows the achievable sum DoF of the proposed caching scheme. From this corollary, we arrive at the following observations:

Observation 4: If $\alpha \in \big(0, \frac{M}{N+1}\big]$, $\underline{d}_{\max} = N$, which is the minimum $\underline{d}_{\max}$ achieved by the proposed wireless caching network. In this case, each user has sufficient spatial dimensions as the number of receive antennas rockets up, leading to a sufficiently large number of bins. As such, the BSs can transmit L NC caching messages as long as $L \leq N$. ■

Consider the symmetric setup with $K = M$ and $N_T = N_R$ without coded prefetching, where each BS pre-downloads the messages without any linear wireless network coding and deploys the precoder based on interference alignment. The sum total is $N_TK/2$ for infinite symbol-extension, and could be far less than $N_TK/2$ for small symbol-extension in practice [25]. The proposed caching network can achieve a higher sum DoF than the uncoded caching.

Remark 2: The underlying delivery network for the proposed wireless caching is a typical distributed MIMO system. The DoF upper bound for the distributed MIMO system is $\underline{d}_{\text{upper}} = \min\{N_TK, N_RM\}$ [35]. Let us validate the achievability of the upper bound by the proposed scheme. On one hand, consider that $N_TK \geq N_RM$. In this case, $\underline{d}_{\text{upper}} = N_RM$. Theorem 3 verifies that the achievable DoF of the proposed scheme yields $\underline{d} \leq \underline{d}_{\max} = \max L$, and $L \leq N_RM$ according to the statement before (C1)-(C3). Following Corollary 4, $\underline{d}_{\max} = \underline{d}_{\text{upper}} = N_RM$ if $\alpha \in \big(M - \frac{N}{N_R}, M - \frac{N}{N_R+1/M}\big]$. On the other hand, consider that $N_TK < N_RM$. In this case, $\underline{d}_{\text{upper}} = N_TK$. Following Corollary 4, $\underline{d}_{\max} = \underline{d}_{\text{upper}} = N_TK$ if $\alpha \in \big(M(1-\frac{N}{N_TK}), M(1-\frac{N}{N_TK+1})\big]$. In this context, one can not only characterize the achievable DoF under arbitrary caching network parameters such as $\{N_T, K, N_R, M, L\}$, but also configurate the caching network to meet the target DoF as long as the target DoF is no larger than the upper bound. ■

Example 1 (Part VI: achievable $\underline{d}_{\max}$) Consider that $K = M = 2$ and $N_T = N_R = 3$. Let $N = \max\{L_1, L_2\} = 3$. In the worst-case caching, $\underline{d}_{\max} = 5$ according to Observation 4. This achievable $\underline{d}_{\max}$ is consistent with the analytical results in Parts I-V of Example 1. In this case, the work in [19] achieves $\underline{d}_{\max} = 4$ (see (39) in [19]). ■

Example 2 (Part IV: achievable $\underline{d}_{\max}$) Consider that $K = M = 3$, $N_R = 3$ and $\underline{d} = 7$. Let $N = \max\{L_1, L_2, L_3\} = 3$. By Theorem 2, it is required that $N_T = 6$, as illustrated in Parts I-III of Example 2. Furthermore, given $K = M = 3$, $N_T = 6, N_R = 3$, and $N = 3$, the proposed scheme can achieve $\underline{d}_{\max} = 8$ by Theorem 3. One can easily formulate the NC caching matrix and SSA from the previous sections. In this case, the work in [19] achieves $\underline{d}_{\max} = 54/7 < 8$ (see d_{11} in [19]). In addition, we remark that [19] can also achieve a higher DoF $\underline{d}_{\max} = 9 > 8$ (see d_{02} in [19]) by interference neutralization, allowing different BSs to share the common caching messages. However, the proposed scheme does not rely on the caching share among different BSs, as explained in Remark 1. ■

According to Theorem 3 and Corollary 3, we arrive at the following upper bounds for achievable rates:

Corollary 5: Consider that user m requests $\mathbf{w}_m = [w_{m,1}, \ldots, w_{m,L_m}]^{\mathcal{T}}$. For the proposed caching scheme, each relay can recover its NC caching message vector if $L \leq N + \lceil \frac{N_T N}{M N_R - N_T} \rceil - 1$ in Corollary 3 and the underlying message rate R_m of $\mathbf{w}_m$ satisfies

$$R_m \leq \sum_{l=1}^{L_m} \min_{\{k,l_k\}: g_{k,l_k}[I(m,l)] \neq 0} R_{k,l_k}^{\mathrm{NC}}, \quad l_k \in \{1, 2, \ldots, L_k\}, k \in \{1, 2, \ldots, K\}, \tag{33}$$

where R_{k,l_k}^{NC} denotes the achievable rate of NC caching message u_{k,l_k} yielding

$$R_{k,l_k}^{\mathrm{NC}} < \min_{m \in \{1,\ldots,M\}} \log_2(\rho |\mathbf{q}_{m,l_m} \mathbf{H}_{k,m} \mathbf{p}_{k,l_k}|^2) \log_2 q, \quad I(k, l_k) \in \mathcal{J}(\mathbf{b}_{m,l_m}), \tag{34}$$

$g[\cdot]$ was defined in (4), $\mathbf{q}_{m,l_m}$ in (30), $\mathcal{J}(\cdot)$ in (10), and $\mathbf{p}_{k,l_k}$ associated with the precoding matrix in Section IV-B.

Proof: As stated in Section III-B, L NC caching messages are encoded by the nested lattice coding design in [24], where these NC caching messages correspond to L message sequences that are encoded by L fine nested lattices into L length-n codewords. In this context, the received signal in (6) can be treated as that induced by transmission of a coded and modulated message over a single channel use. As elaborated in Section IV-D, user m decouples the bins by the receiver shaping in (30), and carries out the reverse NC operation to obtain multiple superimposed signals each containing a linear combination of multiple NC caching messages. Suppose that u_{k,l_k} is involved in user m after the receiver shaping, i.e., $I(k, l_k) \in \mathcal{J}(\mathbf{b}_{m,l_m})$. Finally, each user employs the lattice decoding to decode its requested messages according to a common coarse lattice. With reference to [24], there exists a chain of nested lattice codes such that the decoding error probability of u_{k,l_k} at user m vanishes as $n \to \infty$ if $R_{k,l_k}^{\mathrm{NC},m} < \log_2(\rho |\mathbf{q}_{m,l_m} \mathbf{H}_{k,m} \mathbf{p}_{k,l_k}|^2) \log_2 q$, where $\rho |\mathbf{q}_{m,l_m} \mathbf{H}_{k,m} \mathbf{p}_{k,l_k}|^2$ denotes the effective SNR. Taking the minimum of $R_{k,l_k}^{\mathrm{NC},1}, \ldots, R_{k,l_k}^{\mathrm{NC},M}$ over all M users, the NC caching message of u_{k,l_k} can be reliably computed if $R_{k,l_k}^{\mathrm{NC}} < \min_{m \in \{1,\ldots,M\}} R_{k,l_k}^{\mathrm{NC},m}$. Here we borrow the computation rate in Theorem 3 of [24]. However, different from the rate expression in Theorem 3 of [24], the achievable rate of each NC caching message in (34) only involves the effective SNR of u_{k,l_k} at user m, i.e., $\rho |\mathbf{q}_{m,l_m} \mathbf{H}_{k,m} \mathbf{p}_{k,l_k}|^2$. This is due to the fact that, by the SSA in Section IV, the bin that contains u_{k,l_k} is free of interference from other bins at user m after the receiver shaping in Section IV-D, and the spatial streams that fall inside the same bin as u_{k,l_k} have the identical signal space according to (16).

According to the statement of this theorem, user m requests $\mathbf{w}_m = [w_{m,1}, \ldots, w_{m,L_m}]^{\mathcal{T}}$. W.l.o.g., let us consider w_{m,l_m}. In this case, the indices of NC caching messages $u_{k,l_k}, k \in \{1, \ldots, K\}, l_k \in \{1, \ldots, L_k\}$ that involve w_{m,l_m} yield $\{k, l_k\} : g_{k,l_k}[I(m, l_m)] \neq 0$. From [24], the message rate of w_{m,l_m} yields $R_{m,l_m} \leq \min_{\{k,l_k\}: g_{k,l_k}[I(m,l_m)] \neq 0} R_{k,l_k}^{\mathrm{NC}}$. Summing up over $l_m = 1, \ldots, L_m$, the achievable rate of $\mathbf{w}_m$ is given in (33). ∎

VI. Caching at Users for Rate Constrained Fronthaul

In the wireless caching networks, user devices (such as mobile phones and laptops) are also installed with the memory units to prefetch requested contents in the placement phase [2], [16]–[19]. For the rate constrained fronthaul, BSs cannot prefecth enough caching messages to serve the requesting users, and each user cannot prefetch all of its requested messages due to limited cache size [7]. In this case, caching at users along with insufficient caching at BSs can fulfill the same user requests as the worst-case caching.

A. Cache Placement Over Rate Constrained Fronthaul

Let $\underline{C}$ denote the maximum rate supported by the fronthaul from central unit to the BSs. Therefore, in the cache placement phase, all BSs connected to the central unit can successfully prefetch the NC caching messages (i.e., the error-free placement) through the fronthaul if $\sum_{k=1}^{K} H(\mathbf{u}_k) \leq \underline{C}$, where $H(\mathbf{u}_k)$ denotes the information entropy of $\mathbf{u}_k$. As in [25], this paper considers that the capacity of air-interface is on the same order as that of the fronthaul. Along this line, the fronthaul capacity is described as $\underline{C} = \underline{L} \log \rho + \epsilon$, where $\underline{L}$ is the maximum number of independent data streams of the fronthaul. Specifically, $\underline{L}$ is comparable to the DoF of air interface.

Insufficient Caching at BSs: Consider that the user request is $\{\gamma_1, \ldots, \gamma_M\}$ with a sum DoF $\underline{d} = L$. To achieve $\underline{d} = L$, by Theorem 3, each BS k needs to prefetch the L_k NC caching messages such that $L = L_1 + L_2 + \ldots + L_K$ and each involves the linear combination of $\mathbf{w}_{1\sim M}$, referred to the sufficient caching. Now, consider that the rate constraint on the fronthaul is $\underline{L} \leq L$. This implies that some BSs fail to prefetch enough NC caching messages to support $\underline{d} = L$. Suppose that K' BSs cannot prefetch enough NC caching messages, denoted by BSs $\{1, 2, \ldots, K'\}$ with $K' \leq K$, which are called *underloaded* BSs. Under the rate constraint, the NC caching message vector prefetched by each BS k is designed as

$$\mathbf{u}_k = [u_{k,1}, u_{k,2}, \ldots, u_{k,\tilde{L}_k}]^{\mathcal{T}} = \mathbf{G}_k \otimes \mathbf{w}'_{1\sim M}, \tag{35}$$

where $\mathbf{w}'_{1\sim M} = [\mathbf{w}_1^{\mathcal{T}} \ \ldots \mathbf{w}_{M'}^{\mathcal{T}} \ \mathbf{w}_{M'+1}^{\mathcal{T}} \ldots \mathbf{w}_M^{\mathcal{T}}]^{\mathcal{T}}$. Note that $\mathbf{w}_{m'} = [w_{m',1} \ldots w_{m',\tilde{L}_{m'}}]^{\mathcal{T}}$ and $\tilde{L}_{m'} \leq L_{m'}$, $m' = 1, \ldots, M'$, while $\mathbf{w}_m = [w_{m,1} \ldots w_{m,L_m}]^{\mathcal{T}}$, $m = M' + 1, \ldots, M$. From (35), there exist some underloaded BSs $k' = 1, 2, \ldots, K'$ each with length of NC caching message vector yielding $\tilde{L}_{k'} \leq L_{k'}$, since $\mathbf{w}'_{1\sim M}$ is a subvector of $\mathbf{w}_{1\sim M}$. Meanwhile, BSs $k = K'+1, \ldots, K$ prefetch sufficient NC caching messages with the same length as the non-constraint case, i.e., $\tilde{L}_k = L_k$. Due to the rate constraint, $\sum_{k'=1}^{K'} \tilde{L}_{k'} + \sum_{k=K'+1}^{K} L_k = \underline{L} \leq L$.

Caching at Users: In this case, some user requests cannot be guaranteed due to the insufficient caching messages delivered from BSs. To address this problem, we allow the users to store

[7]It is possible to predict future requests of users by smartly exploiting the statistical traffic patterns and user context information [36]. Moreover, [36] shows that the peak load of cellular networks can be minimized by proactively pre-caching desired information to selected users before they actually request it.

a subset of their requested messages in their local caches. Consider the worst-case caching scenario $\gamma_m = m, \forall m \in \{1,\ldots,M\}$. Suppose that only users $\{1,2,\ldots,M'\}$ prefetch the messages from the central unit as

$$\mathbf{v}_{m'} = [w_{\gamma_{m'},\tilde{L}_{m'}+1}, w_{\gamma_{m'},\tilde{L}_{m'}+2},\ldots,w_{\gamma_{m'},L_{m'}}]^{\mathcal{T}}, \quad m' \in \{1,2,\ldots,M'\}. \tag{36}$$

Following the proposed delivery strategy, the BSs deliver the remaining requested messages that are not included in the caching messages of all users.

B. Content Delivery Phase With User Caching

Considering the K' underloaded BSs with the insufficient caching messages, the received signal vector at each user in the content delivery follows (6). The main differences lie in the NC caching messages broadcasted by the underloaded BSs and precoding matrices of these BSs.

Precoding Designs: We first go through the SSA enabled bin design at users. Given the joint bin matrix $\mathbf{B}_{1\sim M}$ in Theorem 1, we only need to select the columns of $\mathbf{B}_{1\sim M}$ corresponding to the insufficient NC caching messages at all BSs. Obviously, the new bin matrix $\mathbf{B}'_{1\sim M}$ still complies with the design criteria of (C1)-(C3). From (C2), the $\mathbf{B}'_{1\sim M}$ differs from $\mathbf{B}_{1\sim M}$ in the number of bins, since the maximum number induced by the insufficient caching may be decreased. Correspondingly, we can examine the existence of the joint precoding matrix $\mathbf{P}'_{1\sim K}$ for the rate constrained fronthaul by Theorem 2.

After the receiver shaping and reverse NC operation, users m' can only retrieve $\tilde{L}_{m'}$ messages, $m' = 1,2,\ldots M'$, while the request of the remaining users $M'+1,\ldots,M$ are fulfilled in the same way as the sufficient case. This means that K BSs can only provide the DoF of $\tilde{L}_{m'}$ for users m' and the achievable $\underline{d} = \tilde{L}_1 + \ldots + \tilde{L}_{M'} + L_{M'+1} + \ldots + L_M \le L$. Note that user m' has already pre-downloaded the remaining $L_{m'} - \tilde{L}_{m'}$ messages directly from the central unit in (36). Therefore, the user request can be fulfilled by the proposed caching at users and underloaded BSs, and the same $\underline{d} = L$ can be achieved as the scheme without the rate constraint.

It is worthwhile stressing that this subsection follows the same antenna configuration as the caching scenario without the rate constraint. In this context, there exist other solutions of bin matrices and thereby the precoding matrices, since more dimension are left at each user to align the desired messages as the cache size of some BSs is reduced. To be consistent, we construct the bin matrices following Section IV-A.

Example 4 (Insufficient Caching): Consider a rate constraint with $\underline{L} = 4$. This example follows the network configuration of Example 1. Suppose that $\{\gamma_1,\gamma_2\} = \{1,2\}$.

Under the rate constraint, BS 1 is underloaded with an insufficient caching of $\mathbf{u}_1 = [u_{1,1}\ u_{1,2}]^{\mathcal{T}} = \mathbf{G}_1 \otimes \mathbf{w}'_{1\sim 2}$, while BS 2 sufficiently stores $\mathbf{u}_2 = [u_{2,1}\ u_{2,2}]^{\mathcal{T}} = \mathbf{G}_2 \otimes \mathbf{w}'_{1\sim 2}$, where $\mathbf{w}'_{1\sim 2} = [w_{1,1}\ w_{1,2}\ w_{2,1}\ w_{2,2}]^{\mathcal{T}}$. Then, BSs 1 and 2 can reliably pre-download these caching messages, since $L_1 + L_2 = 4 \le \underline{L}$. The caching at user 1 is $\mathbf{v}_1 = w_{1,3}$. According to the design criteria of (C1)-(C3), the joint bin matrix is designed as

$$\mathbf{B}'_{1\sim 2} = \left[\begin{array}{c} \mathbf{B}'_1 \\ \mathbf{B}'_2 \end{array}\right] = \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{array}\right], \tag{37}$$

where $\mathbf{B}'_{1\sim 2}$ selects the $\{1,2,4,5\}$th columns in $\mathbf{B}_{1\sim 2}$ of (12). Note that $\mathbf{P}_1 = [\mathbf{p}_{1,1}\ \mathbf{p}_{1,2}]$ in this example. Given $\mathbf{B}'_{1\sim 2}$, the joint precoding precoding constraint matrix is given by

$$\mathbf{\Gamma}'_{1\sim 2} = \left[\begin{array}{c} \mathbf{\Gamma}_1 \\ \mathbf{\Gamma}_2 \end{array}\right] = \left[\begin{array}{cccc} \mathbf{0} & \mathbf{H}_{1,1} & \mathbf{H}_{2,1} & \mathbf{0} \\ \hline \mathbf{H}_{1,2} & \mathbf{0} & \mathbf{H}_{2,2} & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_{1,2} & \mathbf{0} & \mathbf{H}_{2,2} \end{array}\right]. \tag{38}$$

Correspondingly, the SSA in bin 2 of user 1 yields $\mathbf{H}_{1,1}\mathbf{p}_2 \| \mathbf{H}_{2,1}\mathbf{p}_3$ (i.e., $\mathbf{H}_{1,1}\mathbf{p}_{1,2} \| \mathbf{H}_{2,1}\mathbf{p}_{2,1}$), while each of bins 1 and 3 at user 1 contains a single signal vector. In addition, the SSA in bin 1 of user 2 is $\mathbf{H}_{1,2}\mathbf{p}_1 \| \mathbf{H}_{2,2}\mathbf{p}_3$ (i.e., $\mathbf{H}_{1,2}\mathbf{p}_{1,1} \| \mathbf{H}_{2,2}\mathbf{p}_{2,1}$), and the SSA in bin 2 of user 2 is $\mathbf{H}_{1,2}\mathbf{p}_2 \| \mathbf{H}_{2,2}\mathbf{p}_4$ (i.e., $\mathbf{H}_{1,2}\mathbf{p}_{1,2} \| \mathbf{H}_{2,2}\mathbf{p}_{2,2}$). Therefore, by Theorem 2, we can find at least one precoding matrix $\mathbf{P}'_{1\sim 2}$, since the size of $\mathbf{\Gamma}'_{1\sim 2}$ is 9×12.

By the design criteria of (D1)-(D3), the inverse NC matrix is given by

$$\tilde{\mathbf{G}}_{1\sim 2} = \left[\begin{array}{c} \tilde{\mathbf{G}}_1 \\ \tilde{\mathbf{G}}_2 \end{array}\right] = \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{array}\right]. \tag{39}$$

As an inverse matrix of $\tilde{\mathbf{G}}_{1\sim 2}$ over $\mathbb{F}_q$, the joint NC caching matrix is given by

$$\mathbf{G}_{1\sim 2} = \left[\begin{array}{c} \mathbf{G}_1 \\ \mathbf{G}_2 \end{array}\right] = \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ \hline 1 & 1 & q-1 & 0 \\ q-1 & 0 & 1 & 0 \\ q-1 & q-1 & 1 & 1 \end{array}\right]. \tag{40}$$

After the receiver shaping and reverse NC operation, user 1 can decode $\{w_{1,1}, w_{1,2}\}$, while user 2 can recover $\mathbf{w}_2$. Combined with $w_{1,3}$ at its cache, user 1 can recover $\mathbf{w}_1$.

Remark 3: Ref. [33] proposed an information-theoretic model of a fog-aided system with fronthaul rate limitation, consisting of a cloud, M edge nodes, and K users. The zero-forcing precoding and interference alignment are employed in the delivery phase. According to Lemmas 2 and 3 in [33], the achievable DoF for any fronthaul rate constraint is either $KM/(M+K-1)$ under a fractional cache size $\mu = 1/M$ by the interference alignment, or $\min\{M,K\}$ under $\mu = 1$ (all files cached in each edge node) by the zero-forcing precoding. For example, [33] can achieve the sum DoF of either $4/3$ by the interference alignment or 2 by the zero-forcing precoding under $K = M = 2$. As Section VI elaborates, the proposed caching scheme can achieve the sum DoF $\underline{d}$ if the rate constraint yields $d \le \underline{d}_{\max}$. For example, the proposed caching scheme can achieve $\underline{d}_{\max} = 5$ under $K = M = 2$ (see Part VI of Example 1) and $\underline{d}_{\max} = 8$ under $K = M = 3$ (see Part IV of Example 2).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCOMM.2020.3035607, IEEE Transactions on Communications

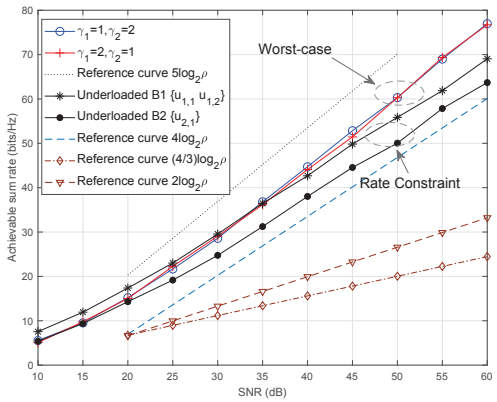


Fig. 5. Achievable sum rate of the proposed wireless caching scheme with $K = M = 2$ and $N_T = N_R = 3$.

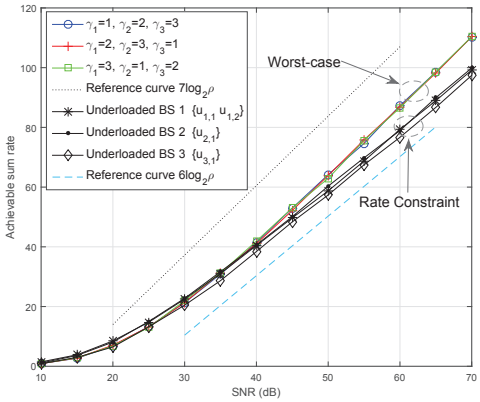


Fig. 6. Achievable sum rate of the proposed wireless caching scheme with $K = M = 3$ and $N_T = 6, N_R = 3$.

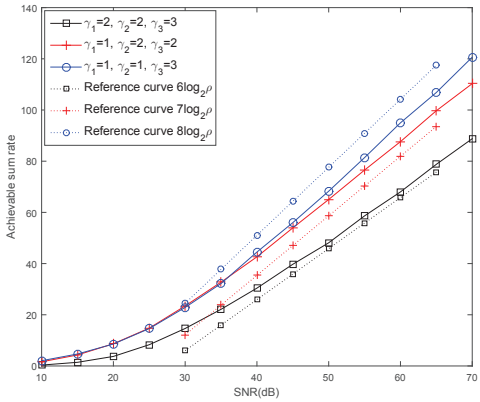


Fig. 7. Achievable sum rate of the proposed wireless caching scheme with $K = M = 3$ and $N_T = 6, N_R = 3$.

VII. Numerical Results

In the section, we provide numerical results to corroborate the analytical results on the achievable sum DoF $\underline{d}$ for different linear NC wireless caching scenarios. To assess the sum DoF, we adopt the achievable sum rate analysis studied in [24], [25]. We note that the sum DoF corresponds to the scaling factor of the sum rate as the SNR goes high.

Fig. 5 plots the achievable sum rates of the proposed 2×2 caching scheme with $N_T = N_R = 3$ under Rayleigh fading channels. We also plot reference rate curves scaling with $\underline{d} = 4$ and $\underline{d} = 5$ respectively. According to Part VI of Example 1, the reference curve with $\underline{d} = 4$ denotes the achievable $\underline{d}_{\max}$ in [19]. From Remark 3, the reference curves with $\underline{d} = 4/3$ and $\underline{d} = 2$ correspond to the achievable DoFs in [33] by the interference alignment and ZF precoding respectively. First, we plot the rate curves for the worst-case caching with different user requests $\{\gamma_1, \gamma_2\} = \{1, 2\}$ and $\{\gamma_1, \gamma_2\} = \{2, 1\}$ (see Example 1). It is observed that the proposed scheme under different user requests can achieve the sum DoF of 5, consistent with the DoF analysis in Section V. In addition, it is shown that the achievable sum rates of the proposed scheme under different $\{\gamma_1, \gamma_2\}$ are close to each other. This result further verifies that the joint NC caching design is compatible with variability of user requests. Second, we plot the rate curves for the caching under the rate constraint fronthaul with $L^o = 4$. To demonstrate Example 4, we observe that the insufficient caching at BS 1 can achieve the sum DoF of 4. Following this example, we also plot the rate curve of the caching with underloaded BS 2 and sufficient caching at BS 1. In this case, the proposed caching scheme can achieve the sum DoF of 4. Combined with the caching at users, the proposed scheme can achieve a sum DoF of 5, consistent with the analytical results in Section VI.

Fig. 6 plots the achievable sum rates of the proposed 3×3 caching scheme with $N_T = 6, N_R = 3$ under rayleigh fading channels. Reference rate curves scaling with $\underline{d} = 6$ and $\underline{d} = 7$ are plotted respectively. First, we demonstrate the sum DoF for the worst-case caching with different user requests $\{\gamma_1, \gamma_2, \gamma_3\} = \{1, 2, 3\}$ (see Example 2), $\{2, 3, 1\}$, and $\{3, 1, 2\}$. We observed that the proposed scheme under different user requests can achieve $\underline{d} = 7$ and the rate curves under different requests are close to each other. Second, we plot the rate curves for the caching under the rate constraint fronthaul with $L^o = 6$. Specifically, we consider insufficient caching at BS 1, 2, and 3 to meet the rate constraint respectively. It is observed that the proposed insufficient caching can achieve the maximum $\underline{d} = 6$ under the rate constraint.

Fig. 7 plots the achievable sum rates of the proposed 3×3 caching scheme with overlapping user requests. Reference rate curves scaling with $\underline{d} = 6, 7, 8$ are plotted respectively. First, we observe that the proposed caching scheme for the overlapping requests $\{\gamma_1, \gamma_2, \gamma_3\} = \{1, 2, 2\}$ can achieve $\underline{d} = 7$, consistent with the analytical results in Example 3. Second, it is shown that the proposed caching scheme for $\{\gamma_1, \gamma_2, \gamma_3\} = \{1, 1, 3\}$ and $\{2, 2, 3\}$ can achieve the sum DoF of 8 and 6 respectively. This agrees with the DoF analysis in Section V.

VIII. Conclusion

We have proposed the linear wireless network coding operated caching network. In the cache placement phase, each BS stores the NC caching messages as a form of linear wireless network coding. In the content delivery phase, we designed the SSA enabled precoding matrix to align the desired NC caching messages based on the inverse NC matrix. With the receiver shaping and reverse NC operation, the proposed scheme can deal with distinct user requests using an invariant cache placement, without the common caching messages at

different BSs. Furthermore, the worst-case caching strategy is amendable to the insufficient caching under a rate constraint fronthaul. Analytical and numerical results showed that the proposed scheme achieves a higher sum DoF than the existing related work.

Several interesting directions follow this work. First, this paper shows that the linear NC wireless caching is applicable in C-RAN. It is of interest to generalize the spirit of linear NC wireless caching into other wireless networks. Second, consider that each cache-aided user uniformly pre-downloads the subfiles of each file in the placement phase. To improve the DoF under limited cache memory, a joint design of linear wireless wireless network coded caching at the BSs and users deserves further investigation. Third, normalized delivery time is used to evaluate the high-SNR latency performance metric in the wireless caching network. Taking different fronthaul-edge transmission models into consideration, the characterization of this metric for the proposed caching scheme deserves our future study.

APPENDIX

W.l.o.g. suppose that $L_{\gamma_1} = \max_{\forall \gamma_m \in \{1,\ldots,K\}} L_{\gamma_m}$. Then, $N = L_{\gamma_1}$.

Case 1: Consider that $L_{\gamma_1} > L_{\gamma_m}$, $\forall m \in \{2,\ldots,M\}$. W.l.o.g., suppose that $L_{\gamma_1} > L_{\gamma_2} \geq \ldots \geq L_{\gamma_K}$. The joint bin matrix $\mathbf{B}_{1\sim M} \in \mathbb{F}_2^{ML_1 \times L}$ is designed as a block-wise manner:

$$\mathbf{B}_{1\sim M} = \begin{bmatrix} \mathbf{B}_{1,1} & \mathbf{B}_{1,2} & \ldots & \mathbf{B}_{1,M} \\ \mathbf{B}_{2,1} & \mathbf{B}_{2,2} & \ldots & \mathbf{B}_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{B}_{M,1} & \mathbf{B}_{M,2} & \ldots & \mathbf{B}_{M,M} \end{bmatrix}, \quad (41)$$

where $\mathbf{B}_{m,m'} \in \mathbb{F}_2^{L_{\gamma_1} \times L_{\gamma_{m'}}}$ denotes the (m,m')th submatrix of $\mathbf{B}_{1\sim M}$, $m,m' \in \{1,2,\ldots,M\}$. To satisfy (C1), each submatrix is given by

$$\mathbf{B}_{m,m} = \begin{bmatrix} \mathbf{I}_{L_{\gamma_m}} \\ \mathbf{0}_{(L_1 - L_{\gamma_m}) \times L_{\gamma_m}} \end{bmatrix}, m = 1,2,\ldots,M, \quad (42)$$

$$\mathbf{B}_{m',m} = \begin{bmatrix} \mathbf{0}_{1 \times L_{\gamma_m}}, \\ \mathbf{I}_{L_m} \\ \mathbf{0}_{(L_1 - L_{\gamma_m} - 1) \times L_{\gamma_m}} \end{bmatrix}, m' = 1,2,\ldots,m-1, \quad (43)$$

$$\mathbf{B}_{m'',m} = \mathbf{B}_{m,m},\ m'' = m+1,\ldots,M, \quad (44)$$

where $\mathbf{0}_{a\times b}$ denotes an $a \times b$ zero matrix.

First, it can be verified that the $\mathbf{B}_{1\sim M}$ in (41) owns the properties of (C1) and (C2). Next, we prove that this $\mathbf{B}_{1\sim M}$ satisfies (C3), i.e., $\text{qfrank}(\mathbf{B}_{1\sim M}) = L$ over $\mathbb{F}_2$.

Subtracting the $i-1$th submatrix-row[8] from the ith submatrix-row of $\mathbf{B}_{1\sim M}$ in (41), $i = M, M-1, \ldots, 2$, we can transform this $\mathbf{B}_{1\sim M}$ into a submatrix-row echelon form as follows:

$$\mathbf{B}'_{1\sim M} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{B}_{1,2} & \ldots & \mathbf{B}_{1,M} \\ \mathbf{0} & \mathbf{A}_2 & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{A}_M \end{bmatrix}, \quad (45)$$

[8]By the ith submatrix-row, we mean the ith row formed by the submatrices $\{\mathbf{B}_{i,1},\ldots,\mathbf{B}_{i,M}\}$ in (41).

where $\mathbf{A}_1 = \mathbf{B}_{1,1}$ and $\mathbf{A}_m = \mathbf{B}_{m,m} - \mathbf{B}_{m-1,m} = \begin{bmatrix} \boldsymbol{\Upsilon}_m \\ \mathbf{0}_{(L_{\gamma_1} - L_{\gamma_m} - 1) \times L_{\gamma_m}} \end{bmatrix}$, $m = 2,\ldots,M$, with a $(L_{\gamma_m}+1) \times L_{\gamma_m}$ Toeplitz matrix yielding

$$\boldsymbol{\Upsilon}_m = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ 1 & 1 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 & 1 \\ 0 & 0 & \ldots & 0 & 1 \end{bmatrix}. \quad (46)$$

Note that the Toeplitz matrix in (46) has the full rank of L_{γ_m}. Now, it is evident that $\mathbf{B}'_{1\sim M}$ has the full rank over $\mathbb{F}_2$, since each submatrix in the diagonal has the full rank of L_{γ_m} over $\mathbb{F}_2$.

Therefore, $\mathbf{B}_{1\sim M}$ in (41) has the full rank of L, consistent with (C3).

Case 2: There exists at least one L_{γ_k} such that $L_{\gamma_1} = L_{\gamma_k}$, $\exists k \in \{2,\ldots,K\}$. W.l.o.g., suppose that $L_{\gamma_1} = L_{\gamma_2} \geq \ldots \geq L_{\gamma_K}$. In this case, $N = L_{\gamma_1} = L_{\gamma_2}$. We prove that it is not possible to design a joint bin matrix $\mathbf{B}_{1\sim M}$ satisfying (C1)-(C3). Let us focus on the first $2L_1\gamma_1$ columns of $\mathbf{B}_{1\sim M}$. To fulfill (C1), it can be verified that the 1st $\sim L_1$th columns of $\mathbf{B}_{1\sim M}$ (i.e., the 1st submatrix-column) is linearly dependent on the L_1+1th $\sim 2L_1$th columns $\mathbf{B}_{1\sim M}$ (i.e., the 2nd submatrix-column) over $\mathbb{F}_2$. As a result, $\mathbf{B}_{1\sim M}$ is rank deficient over $\mathbb{F}_2$, contradicting with (C3).

REFERENCES

[1] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. M. 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] D. Liu, B. Chen, C. Yang, and A. F. Molisch, "Caching at the wireless edge: design aspects, challenges, and future direction," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 22–28, Sep. 2016.

[3] G. Paschos, E. Bastug, I. Land, G. Caire, and M. Debbah, "Wireless caching: technical misconceptions and business barriers," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 16–22, Aug. 2016.

[4] J. Li, H. Chen, Y. Chen, Z. Lin, B. Vucetic, and L. Hanzo, "Pricing and resource allocation via game theory for a small-cell video caching system," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 8, pp. 2115–2129, Aug. 2016.

[5] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," *IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 849–869, Feb. 2016.

[6] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: basic principles and system performance," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 1, pp. 176–189, Jan. 2016.

[7] W. Han, A. Liu, and V. K. Lau, "PHY-caching in 5G wireless networks: design and analysis," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 30–36, Aug. 2016.

[8] M. Tao, E. Chen, and W. Yu, "Content-centric sparse multicast beamforming for cache-enabled cloud RAN," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6118–6131, Sept. 2016.

[9] H. Wei, A. Liu, and V. K. Lau. "Degrees of freedom in cached MIMO relay networks," *IEEE Trans. Signal Process.*, vol. 63, no. 15, pp. 3986–3997, Aug. 2015.

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

[11] M. A. Maddah-Ali and U. Niesen, "Coding for caching: fundamental limits and practical challenges," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 23–29, Aug. 2016

[12] J. Zhang and P. Elia, "Wireless coded caching: a topological perspective," in *Proc. IEEE ISIT*, Jun. 2017, pp. 401–405.

[13] A. Tang, S. Roy, and X. Wang, "Coded caching for wireless backhaul networks with unequal link rates," *IEEE Trans. Commun.*, vol. 66, no. 1, pp. 1–13, Jan. 2018.

[14] S. P. Shariatpanahi, G. Caire, and B. H. Khalaj, "Physical-layer schemes for wireless coded caching," *IEEE Trans. Inf. Theory*, vol. 65, no. 5, pp. 2792–2807, May 2019.

[15] Y. Uğur, Z. H. Awan, and A. Sezgin, "Cloud radio access networks with coded caching," in *Proc. WSA*, Mar. 2016, pp. 160–164.

[16] N. Naderializadeh, M. A. M.-Ali, and A. S. Avestimehr, "Fundamental limits of cache-aided interference management," *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 3092–3107, May 2017.

[17] J. Hachem, U. Niesen, and S. Diggavi, "Degrees of freedom of cache-aided wireless interference networks", *IEEE Trans. Inf. Theory*, vol. 64, no. 7, pp. 5359–5380, Apr. 2018.

[18] F. Xu, M. Tao, and K. Liu, "Fundamental tradeoff between storage and latency in cache-aided wireless interference networks", *IEEE Trans. Inf. Theory*, vol. 63, no. 11, pp. 7464–7491, Jun. 2017.

[19] Y. Cao, M. Tao, F. Xu, and K. Liu, "Fundamental storage-latency tradeoff in cache-aided MIMO interference networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 5061–5076, Aug. 2017.

[20] S. Zhang, S. C. Liew, and P. P. Lam, "Hot topic: physical-layer network coding," in *Proc. ACM Mobicom*, Sept. 2006.

[21] T. Yang and I. Collings, "On the optimal design and performance of linear physical-layer network coding for fading two-way relay channels," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 956–967, Feb. 2014.

[22] L. Shi, S. C. Liew, and L. Lu, "On the subtleties of q-PAM linear physical-layer network coding," *IEEE Trans. Inf. Theory*, vol. 63, no. 8, pp. 2520–2544, May 2016.

[23] L. Shi and S. C. Liew, "Complex linear physical-layer network coding," *IEEE Trans. Inf. Theory*, vol. 62, no. 5, pp. 4949–4981, Aug. 2017.

[24] B. Nazer and M. Gastpar, "Compute-and-forward: harnessing interference through structured codes," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6463–6485, Oct. 2011.

[25] T. Yang, "Distributed MIMO broadcasting: reverse compute-and-forward and signal space alignment," *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 581–593, Jan. 2017.

[26] S.-N. Hong and G. Caire, "Compute-and-forward strategies for cooperative distributed antenna systems," *IEEE Trans. Inf. Theory*, vol. 59, no. 9, pp. 5227–5243, Sep. 2013.

[27] V. R. Cadambe and S. A. Jafar, "Interference alignment and the degrees of freedom of the K user interference channel", *IEEE Trans. Inf. Theory*, vol. 54, no. 8, pp. 3425–3441, Aug. 2008.

[28] N. Lee, J.-B. Lim, and J. Chun, "Degrees of freedom of the MIMO Y channel: signal space alignment for network coding," *IEEE Trans. Inf. Theory*, vol. 56, no. 7, pp. 3332–3342, Jul. 2010.

[29] *Rank-nullity theorem*, https://en.wikipedia.org/wiki/Rank-nullity_theorem

[30] L. Shi, K. Cai, T. Yang, and T. Wang, "Linear network coded wireless caching," in *Proc. IEEE ICC-ICECC Workshop*, May 2018.

[31] L. Yang, T. Yang, J. Yuan, and J. An, "Achieving the near-capacity of two-way relay channels with modulation-coded physical-layer network coding," *IEEE Trans. Wireless Commun.*, vol. 14, no. 9, pp. 5225–5239, May 2016.

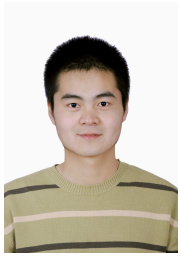
[32] B. Dai, Y.-F. Liu, and W. Yu, "Optimized base-station cache allocation for cloud radio access network with multicast backhaul,", *IEEE J. Sel. Areas Commun.*, vol. 36, no. 8, pp. 1737–1750, Aug. 2018.

[33] A. Sengupta, R. Tandon, and O. Simeone, "Fog-aided wireless networks for content delivery: fundamental latency tradeoffs," *IEEE Trans. Inf. Theory*, vol. 59, no. 9, pp. 5227–5243, Sep. 2013.

[34] J. Zhang and O. Simeone, "Fundamental limits of cloud and cache-aided interference management with multi-antenna edge nodes," *IEEE Trans. Inf. Theory*, vol. 65, no. 8, pp. 5197–5214, Aug. 2013.

[35] C. S. Vaze and M. K. Varanasi, "The degree-of-freedom regions of MIMO broadcast, interference, and cognitive radio channels with no CSIT," *IEEE Trans. Inf. Theory*, vol. 58, no. 8, pp. 5354–5374, Aug. 2012.

[36] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.

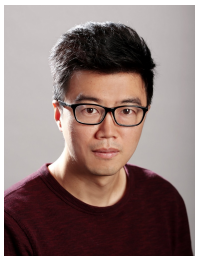


Long Shi (Member, IEEE) received the Ph.D. degree in Electrical Engineering from the University of New South Wales, Sydney, Australia, in 2012. From 2013 to 2016, he was a Postdoctoral Fellow at the Institute of Network Coding, Chinese University of Hong Kong, China. From 2014 to 2017, he was a Lecturer at Nanjing University of Aeronautics and Astronautics, Nanjing, China. From 2017 to 2020, he was a Research Fellow at the Singapore University of Technology and Design. Now he is a Professor at the School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing, China.

Kui Cai (Senior Member, IEEE) received her B.E. degree in information and control engineering from Shanghai Jiao Tong University, Shanghai, China, and joint Ph.D. degree in electrical engineering from Technical University of Eindhoven, The Netherlands, and National University of Singapore. Currently, she is an Associate Professor with Singapore University of Technology and Design (SUTD). She received 2008 IEEE Communications Society Best Paper Award in Coding and Signal Processing for Data Storage. She is an IEEE senior member, and served as the Vice-Chair (Academia) of IEEE Communications Society, Data Storage Technical Committee (DSTC) during 2015 and 2016. Her main research interests are in the areas of coding theory, information theory, and signal processing for various data storage systems and digital communications.

Tao Yang (Member, IEEE) received the B.Sc. degree in electronic engineering from the Beijing University of Aeronautics and Astronautics (Beihang University), China, in 2003, and Ph.D. degrees in electrical engineering from the University of New South Wales (UNSW), Sydney, NSW, Australia, in 2010. He was an OCE Postdoctoral Fellow Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia. He was a Lecturer within the School of Computing and Communications, University of Technology Sydney. He is currently with the Beijing University of Aeronautics and Astronautics (Beihang University). He has authored over 70 research articles in the IEEE journals and conferences. His research expertise and interests include multiple access techniques, network coding, MIMO, error-control coding, iterative signal processing. He has been serving as a TPC member of the IEEE ICC. He was a recipient of the Australian Postgraduate Award, the NICTA Research Project Award. He holds an Australian Research Council Discovery Early Career Research Award Fellowship.

Taotao Wang (Member, IEEE) received the B.S. degree in Electrical Engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2008, the M.S. degree in Information and Signal Processing from the Beijing University of Posts and Telecommunications, Beijing, China, in 2011, and the Ph.D. degree in Information Engineering from the Chinese University of Hong Kong, Hong Kong, in 2015. From 2015 to 2016, he was a Postdoctoral Research Fellow at the Institute of Network Coding, the Chinese University of Hong Kong. After that, he joined the College of Information Engineering at Shenzhen University as an Assistant Professor. His main research interests include distributed network and blockchain technology, wireless communications and networking, statistical signal and data processing. He is a recipient of the Hong Kong Ph.D. Fellowship.



Jun Li (Senior Member, IEEE) received the B.E. and Ph.D. degrees in electronics engineering from the Beijing Institute of Technology, Beijing, China, in 1991 and 1997, respectively. From 1997 to 1999, he was a Research Fellow with the School of Electrical Engineering, University of Sydney, Sydney, Australia. In 2000, he joined the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, Australia, where he is currently a Professor and Head of Telecommunication Group with the School. He has published two books, five book chapters, over 300 papers in telecommunications journals and conference proceedings, and 50 industrial reports. He is a co-inventor of one patent on MIMO systems and two patents on low-density-parity-check codes. He has co-authored four Best Paper Awards and one Best Poster Award, including the Best Paper Award from the IEEE International Conference on Communications, Kansas City, USA, in 2018, the Best Paper Award from IEEE Wireless Communications and Networking Conference, Cancun, Mexico, in 2011, and the Best Paper Award from the IEEE International Symposium on Wireless Communications Systems, Trondheim, Norway, in 2007. He is an IEEE Fellow and currently serving as an Associate Editor for the IEEE Transactions on Wireless Communications. He served as the IEEE NSW Chapter Chair of Joint Communications/Signal Processions/Ocean Engineering Chapter during 2011-2014 and served as an Associate Editor for the IEEE Transactions on Communications during 2012-2017. His current research interests include error control coding and information theory, communication theory, and wireless communications.