Placement delivery array design for combination networks with edge caching

Qifa Yan, Michèle Wigger, Sheng Yang

Placement Delivery Array Design for Combination Networks with Edge Caching

Qifa Yan
LTCI, Télécom ParisTech
75013 Paris, France
Email: qifa.yan@telecom-paristech.fr

Michèle Wigger
LTCI, Télécom ParisTech
75013 Paris, France
Email: michele.wigger@telecom-paristech.fr

Sheng Yang
L2S, CentraleSupélec
91190 Gif-sur-Yvette, France
Email:sheng.yang@centralesupelec.fr

***Abstract*—A major practical limitation of the Maddah-Ali-Niesen coded caching techniques is their high subpacketization level. For the simple network with a single server and multiple users, Yan *et al.* proposed an alternative scheme with the so-called placement delivery arrays (PDA). Such a scheme requires slightly higher transmission rates but significantly reduces the subpacketization level. In this paper, we extend the PDA framework and propose three low-subpacketization schemes for combination networks, i.e., networks with a single server, multiple relays, and multiple cache-aided users that are connected to subsets of relays. One of the schemes achieves the cutset lower bound on the link rate when the cache memories are sufficiently large. Our other two schemes apply only to *resolvable* combination networks. For these networks and for a wide range of cache sizes, the new schemes perform closely to the coded caching schemes that directly apply Maddah-Ali-Niesen scheme while having significantly reduced subpacketization levels.**

I. Introduction

Caching is a promising approach to alleviate current network traffics driven by on-demand video streaming. The idea is to pre-fetch contents during off-peak hours before the actual user demands, so as to reduce traffic at peak hours when the demands are made. Therefore, the communication takes place in two phases: *content placement* at off-peak hours and *content delivery* at peak hours.

In their seminal work [1], Maddah-Ali and Niesen modeled the content delivery phase by a shared error-free link from the single server to all users, and they showed that delivery traffic in this *shared-link setup* can be highly reduced through a joint design of content placement and delivery strategy that exploits multicasting opportunities. The scheme is known as *coded caching* and has been extended to various settings, e.g., Gaussian broadcast channels [2], multi-antenna fading channels [3]–[5], or *combination networks* [6]–[11] as considered in this paper. In a (h, r)-combination network, a single server communicates over dedicated error-free links with h relays and these relays in their turn communicate over dedicated error-free links with $\binom{h}{r}$ users that have local cache memories. Each user is connected to a different subset of r relays. Ji *et al.* first investigated this network [6] for the case when r divides h (denoted by $r|h$), and the achievable bound was improved in [7]. In [8], Wan *et al.* tightened the lower bound under the constraint of uncoded placement, and the achievable bound for the case when the memory size is small. In [10]–[12], Maxmimum Distance Separable (MDS) codes are applied before placement. In particular, [10], [11] show that the upper bound in [7] is achievable for any (h, r) combination network, and [12] shows that even lower delivery rates are achievable. As the results of our work require memory size larger than that of [8] and is uncoded placement, we only compare our results with those from [7].

A key factor that limits the application of all forms of coded caching in practice, is the required high *subpacketization level* [13], i.e., the number of subpackets must grow exponentially with the number of users. In contrast, [14]–[18] proposed new caching schemes that have much lower subpacketization levels but slightly increased transmission rate. A useful tool for representing these new schemes is *placement delivery array (PDA)* introduced in [14]. PDAs characterize both the (uncoded) placement and delivery strategies with a single array [14], and thus facilitate the design of good caching schemes.

In this paper, we first introduce *combinational PDAs (C-PDA)* to represent uncoded placement and delivery strategies for combination networks in a single array. We also determine the rate, memory, and subpacketization requirements of the caching scheme corresponding to a given C-PDA. Then, for the case when $r|h$, we describe how any standard PDA with $\binom{h-1}{r-1}$ columns can be transformed into a C-PDA for a (h, r)-combination network. With this transformation and the previous low-subpacketization schemes for the single-shared link setup, two low-subpacketization schemes for (h, r)-combination networks are obtained. The performances of the new schemes are close to the scheme in [7], but have significantly lower subpacketization level. Finally, for arbitrary (h, r), we propose a C-PDA for which the corresponding caching scheme achieves the cut-set lower bound for sufficiently large cache sizes.

Due to the space limitation, we only provide sketches of the proofs. For details, see [19].

Notations: We denote the set of positive integers by $\mathbb{N}^+$. For $n \in \mathbb{N}^+$, denote the set $\{1, 2, \cdots, n\}$ by $[n]$. The Exclusive OR operation is denoted by $\oplus$. For a positive real number x, $\lceil x \rceil$ is the least integer that is not less than x.

II. System Model and Preliminaries

A. System Model

Consider the (h, r)-combination network illustrated in Fig. 1, where h and r are positive integers and $r \leq h$. The

network comprises of a single server, h relays:

$$\mathcal{H} = \{H_1, H_2, \cdots, H_h\},$$

and $K = \binom{h}{r}$ users labeled by all the r-dimensional subsets of relay indices $[h]$:

$$\mathbf{T} \triangleq \big\{T\colon\ T \subset [h] \text{ and } |T| = r\big\}. \tag{1}$$

Each user has a local cache memory of size MB bits. The relays have no cache memories. The server can directly access a library $\mathcal{W}$ of N files,

$$\mathcal{W} = \{W_1, W_2, \cdots, W_N\},$$

where each file W_n consists of B independent and identically uniformly distributed (i.i.d.) random bits. The server can send RB bits to each of the h relays over an individual error-free link. Here, R is the *link rate* (or *rate* for brevity). Each relay can communicate with some of the users. Specifically, user T is connected through individual error-free links of rate R to the r relays with index in T, i.e., to relays $\{H_i : i \in T\}$.

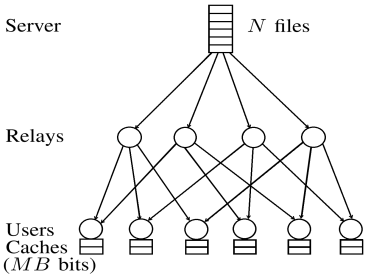

Fig. 1: A $(4, 2)$-combination caching network.

We now describe the storage and communication operations. The system operates in two consecutive phases.

1. Placement Phase: In this phase, each user T directly accesses to the file library $\mathcal{W}$ and can store an arbitrary function thereof in its cache memory, subject to the space limitation of MB bits. Denote the cached content at user T by Z_T, and the set of all cached contents by $\boldsymbol{Z} \triangleq \{Z_T\colon T \in \mathbf{T}\}$.

2. Delivery Phase: In this phase, each user T arbitrarily requests a file W_{d_T} from the server, where $d_T \in [N]$. The users' requests $\boldsymbol{d} \triangleq \{d_T\colon T \in \mathbf{T}\} \in [N]^K$ are revealed to all parties, i.e., to server, relays, and users. For each $i \in [h]$, the server sends RB bits to relay H_i:

$$X_i = \phi_i(W_1, \ldots, W_N, \boldsymbol{Z}, \boldsymbol{d}),$$

for some function $\phi_i\colon \mathbb{F}_2^{B\cdot N} \times \mathbb{F}_2^{B\cdot M\cdot K} \times [N]^K \to \mathbb{F}_2^{B\cdot R}$. Relay H_i forwards the signal X_i to all connected users.[1]

At the end of this phase, each user $T \in \mathbf{T}$, decodes its requested file W_{d_T} based on all its received signals $\boldsymbol{X}_T \triangleq \{X_i\colon i \in T\}$, its cache content Z_T, and demand vector $\boldsymbol{d}$:

$$\hat{W}_{d_T} = \psi_T(\boldsymbol{X}_T, Z_T, \boldsymbol{d}),$$

[1]Previous works on combination networks allow the relay to send different arbitrary functions to their connected users. But since the rate of relay-to-users links needs not exceed the rate of the server-to-relays, this apparently more general setup does not allow for better communication strategies.

for some function $\psi_T\colon \mathbb{F}_2^{B\cdot R\cdot r} \times \mathbb{F}_2^{B\cdot M} \times [N]^K \to \mathbb{F}_2^{B}$.

The *optimal worst-case rate* $R^\star(M)$ is the smallest delivery rate R for which there exist some placement and delivery strategies so that the probability of decoding error $\hat{W}_{d_T} \neq W_{d_T}$ vanishes asymptotically as $B \to \infty$ at all the users and for any possible demand $\boldsymbol{d}$.

Special focus will be given to (h, r)-combination networks with $r|h$. In this case, the users can be partitioned into subsets so that in each subset exactly one user is connected to a given relay, see [7].

Definition 1 (Resolvable Networks). *A combination network is called* resolvable *if the user set* $\mathbf{T}$ *can be partitioned into subsets* $\mathcal{P}_1, \mathcal{P}_2, \cdots, \mathcal{P}_{\tilde{K}}$ *so that for all* $i \in [\tilde{K}]$ *the following two conditions hold:*

- *If* $T, T' \in \mathcal{P}_i$ *and* $T \neq T'$*, then* $T \cap T' = \emptyset$.
- $\bigcup_{T:T\in\mathcal{P}_i} T = [h]$.

Subsets $\mathcal{P}_1, \mathcal{P}_2, \cdots, \mathcal{P}_{\tilde{K}}$ *satisfying these conditions are called* parallel classes.

B. Preliminaries: Shared-link Setup and PDAs

For the purpose of this subsection, consider the original coded caching setup [1] with a single server and K users each having a cache memory of MB bits. The server is connected to the users through a shared error-free link of rate R.

Yan *et al.* [14] proposed to unify the description of uncoded placement and delivery strategies for this shared-link setup in a single array, called the *placement delivery array (PDA)*.

Definition 2 (PDA, [14]). *For positive integers* K, F, Z *and* S*, an* $F \times K$ *array* $\boldsymbol{A} = [a_{j,k}]$*,* $j \in [F], k \in [K]$*, composed of a specific symbol "*$*$*" and* S *ordinary symbols* $1, \cdots, S$*, is called a* (K, F, Z, S) placement delivery array (PDA)*, if it satisfies the following conditions:*

C1. *The symbol "*$*$*" appears* Z *times in each column;*

C2. *Each ordinary symbol occurs at least once in the array;*

C3. *For any two distinct entries* a_{j_1,k_1} *and* a_{j_2,k_2}*, we have* $a_{j_1,k_1} = a_{j_2,k_2} = s$*, an ordinary symbol only if*

   a. $j_1 \neq j_2$*,* $k_1 \neq k_2$*, i.e., they lie in distinct rows and distinct columns; and*

   b. $a_{j_1,k_2} = a_{j_2,k_1} = *$*, i.e., the corresponding* 2×2 *sub-array formed by rows* j_1, j_2 *and columns* k_1, k_2 *must be of the following form*

$$\begin{bmatrix} s & * \\ * & s \end{bmatrix} \text{ or } \begin{bmatrix} * & s \\ s & * \end{bmatrix}. \tag{2}$$

We refer to the parameter F *as the subpacketization level. Specially, if each ordinary symbol* $s \in [S]$ *occurs exactly* g *times,* $\boldsymbol{A}$ *is called a* g*-*(K, F, Z, S) *PDA, or* g*-PDA for short.*

Any PDA can be transformed into a caching scheme having the following performance [14]:

Remark 1. A (K, F, Z, S) PDA corresponds to a caching scheme for the shared error-free link setup with K users that is of subpacketization level F, requires cache size $M = \frac{Z}{F}N$, and delivery rate $R = \frac{S}{F}$.

Two low-subpacketization schemes were proposed in [14]:

Lemma 1 (PDA for $\frac{N}{M} \in \mathbb{N}^+$, [14]). *For any $q, m \in \mathbb{N}^+, q \geq 2$, there exists a $(m+1)$-$(q(m+1), q^m, q^{m-1}, q^{m+1}-q^m)$ PDA, with rate $R = \frac{N}{M} - 1$ and subpacketization level $F = (\frac{N}{M})^{\frac{KM}{N}-1}$.*

Lemma 2 (PDA for $\frac{N}{N-M} \in \mathbb{N}^+$, [14]). *For any $q, m \in \mathbb{N}^+, q \geq 2$, there exists a $(q-1)(m+1)$-$(q(m+1), (q-1)q^m, (q-1)^2 q^{m-1}, q^m)$ PDA, with rate $R = \frac{N}{M} - 1$ and subpacketization level $F = \frac{M}{N-M} \cdot \left(\frac{N}{N-M}\right)^{K(1-\frac{M}{N})-1}$.*

III. C-PDAs for Combination Networks

A PDA is especially useful for a combination network, if for any coded packet, all the intended users are connected to the same relay. This allows the server to send each coded packet only to this single relay. The following definition ensures the desired property.

Definition 3. *Let $h, r \in \mathbb{N}^+$ with $r \leq h$, and $K = \binom{h}{r}$. A (K, F, Z, S) PDA is called* (h, r)-combinational*, for short* C-PDA*, if its columns can be labeled by the sets in $\mathbf{T}$ in a way that for any ordinary symbol $s \in [S]$, the labels of all columns containing symbol s have nonempty intersection.*

The following example presents a $(6, 6, 2, 12)$ C-PDA for $h = 4$ and $r = 2$, and explains how this C-PDA leads to a caching scheme for the $(4, 2)$-combination network in Fig. 1.

Example 1. Let $h = 4$ and $r = 2$. The following table presents a C-PDA combined with a labeling of the columns that satisfies the condition in Definition 3.

TABLE I: A C-PDA for the setting in Fig. 1.

$\{1,2\}$	$\{3,4\}$	$\{1,3\}$	$\{2,4\}$	$\{1,4\}$	$\{2,3\}$
$*$	$*$	1	4	2	5
1	7	$*$	$*$	3	6
2	8	3	6	$*$	$*$
$*$	$*$	7	10	11	8
4	10	$*$	$*$	12	9
5	11	9	12	$*$	$*$

The above C-PDA implies the following caching scheme for the $(h = 4, r = 2)$ combination network in Fig. 1.

1. Placement phase: Each file is split into 6 packets (i.e., the number of rows of the C-PDA), i.e., $W_n = \{W_{n,i} \colon\ i \in [6],\ n \in [N]\}$. Place the following cache contents at the users:

$$Z_{\{1,2\}} = Z_{\{3,4\}} = \{W_{n,1}, W_{n,4} \colon\ n \in [N]\}$$
$$Z_{\{1,3\}} = Z_{\{2,4\}} = \{W_{n,2}, W_{n,5} \colon\ n \in [N]\}$$
$$Z_{\{1,4\}} = Z_{\{2,3\}} = \{W_{n,3}, W_{n,6} \colon\ n \in [N]\}$$

2. Delivery phase: Table II shows the signals $X_1, \ldots, X_4$ the server sends to the four relays when users $U_{\{1,2\}}$, $U_{\{3,4\}}$, $U_{\{1,3\}}$, $U_{\{2,4\}}$, $U_{\{1,4\}}$, $U_{\{2,3\}}$ request files W_1, W_2, W_3, W_4, W_5, W_6, respectively. Each of the coded signals consists of $B/6$ bits, and thus the required rate is $R = 1/2$.

Table II also indicates the users that are actually interested by each coded signal. In the problem definition, we assumed that each relay forwards its entire received signal to all its connected users. From Table II, it is obvious that it would suffice to forward only a subset of the bits to each user.

TABLE II: Delivered signals in Example 1.

Signal	Symbol s	Coded Signal	Intended Users
X_1	1	$W_{1,2} \oplus W_{3,1}$	$U_{\{1,2\}},\ U_{\{1,3\}}$
	2	$W_{1,3} \oplus W_{5,1}$	$U_{\{1,2\}},\ U_{\{1,4\}}$
	3	$W_{3,3} \oplus W_{5,2}$	$U_{\{1,3\}},\ U_{\{1,4\}}$
X_2	4	$W_{1,5} \oplus W_{4,1}$	$U_{\{1,2\}},\ U_{\{2,4\}}$
	5	$W_{1,6} \oplus W_{6,1}$	$U_{\{1,2\}},\ U_{\{2,3\}}$
	6	$W_{4,3} \oplus W_{6,2}$	$U_{\{2,4\}},\ U_{\{2,3\}}$
X_3	7	$W_{2,2} \oplus W_{3,4}$	$U_{\{3,4\}},\ U_{\{1,3\}}$
	8	$W_{2,3} \oplus W_{6,4}$	$U_{\{3,4\}},\ U_{\{2,3\}}$
	9	$W_{3,6} \oplus W_{6,5}$	$U_{\{1,3\}},\ U_{\{2,3\}}$
X_4	10	$W_{2,5} \oplus W_{4,4}$	$U_{\{3,4\}},\ U_{\{2,4\}}$
	11	$W_{2,6} \oplus W_{5,4}$	$U_{\{3,4\}},\ U_{\{1,4\}}$
	12	$W_{4,6} \oplus W_{5,5}$	$U_{\{2,4\}},\ U_{\{1,4\}}$

We now present a general way to associate a (K, F, Z, S) C-PDA to a caching scheme for a (h, r)-combination network where h, r are positive integers with $r \leq h$.

Placement phase: Label the columns of the C-PDA with the set $\mathbf{T}$ so that the condition in Definition 3 is satisfied. Placement is the same as for standard PDAs. That means, split each file W_d into F subpackets $(W_{d,1}, \ldots, W_{d,F})$ each consisting of B/F bits. Place subfiles $\{W_{n,i}\}_{n=1}^{N}$ into the cache memory of user T, if the C-PDA has entry "$*$" in row i and the column corresponding to label T. This placement strategy requires a cache size of $M = N \cdot \frac{Z}{F}$.

Delivery phase: The server first creates the coded signals pertaining to each ordinary symbol $s \in [S]$ in the same way as for standard PDAs. It then delivers the coded signal created for each ordinary symbol $s \in [S]$ to one of the relays whose index is contained in the labels of all columns containing s. The *average* rate required on the h server-to-relay links is $R_{\text{avg}} = \frac{S}{Fh}$.

When in the described scheme the server sends the same number of bits to each relay, then the following theorem follows immediately from the above description. In fact, in this case subpacketization level F is sufficient. Otherwise, the rate on each server-to-relay link has to be made equal by first splitting each file into h subfiles and then applying a caching scheme with the same C-PDA but a different shifted version of the column labels to each of the subfiles.

Theorem 1. *Given a (K, F, Z, S) C-PDA. For any (h, r) combination network with $K = \binom{h}{r}$, it holds that $R^\star\left(M = \frac{N \cdot Z}{F}\right) \leq \frac{S}{Fh}$. This upper bound is achieved by a scheme of subpacketization level not exceeding hF.*

IV. Transforming PDAs into Larger C-PDAs

We present a way of constructing C-PDAs for resolvable (h, r)-combination networks (i.e., when $r|h$) from any smaller PDA that has $\tilde{K} = \binom{h-1}{r-1}$ columns. We start with an example.

Example 2. Reconsider Example 1, where $h = 4$ and $r = 2$, and notice that for this resolvable network (see Definition 1), a possible partition of $\mathbf{T}$ is $\mathcal{P}_1 = \{\{1,2\}, \{3,4\}\}, \mathcal{P}_2 =$

$\{\{1,3\},\{2,4\}\}$ and $\mathcal{P}_3 = \{\{1,4\},\{2,3\}\}$. Consider now the $(3,3,1,3)$ PDA of the Maddah-Ali & Niesen scheme with $\tilde{K}=3$ users:

$$\boldsymbol{A} = \begin{bmatrix} * & 1 & 2 \\ 1 & * & 3 \\ 2 & 3 & * \end{bmatrix}.$$

One can verify that the C-PDA in Table I is obtained from above PDA $\boldsymbol{A}$ by replicating each column of $\boldsymbol{A}$ first horizontally and then each column of the resulting array also vertically, and by then replacing the 3 replicas of each ordinary symbol with 3 new (unused) symbols. The column labels are obtained by labeling the first two columns of $\boldsymbol{A}$ with the two elements of $\mathcal{P}_1$, the following two columns with the elements of $\mathcal{P}_2$, and the last two columns with the elements of $\mathcal{P}_3$.

We now present the general transformation method. We use the following notations. For a given user T, let $\delta(T)$ indicate the parallel class that T belongs to, i.e., $\delta(T)=j$ iff $T\in\mathcal{P}_j$. Let $T[i]$ be the i-th smallest element of T. For example, if $T=\{2,4\}$, then $T[1]=2, T[2]=4$. Likewise, denote the inverse map by T^{-1}, i.e., $T[i]=j$ iff $T^{-1}[j]=i$.

Transformation 1. Given a $(\tilde{K},\tilde{F},\tilde{Z},\tilde{S})$ PDA $\tilde{\boldsymbol{C}}=[\tilde{c}_{j,k}]$. Let the following $(\tilde{F}r)$-by-$(\tilde{K}\frac{h}{r})$ array $\boldsymbol{C}$ be the outcome applied to PDA $\tilde{\boldsymbol{C}}$ for parameters (h,r):

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{c}_{1,T_1} & \boldsymbol{c}_{1,T_2} & \cdots & \boldsymbol{c}_{1,T_K} \\ \boldsymbol{c}_{2,T_1} & \boldsymbol{c}_{1,T_2} & \cdots & \boldsymbol{c}_{2,T_K} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{c}_{r,T_1} & \boldsymbol{c}_{r,T_2} & \cdots & \boldsymbol{c}_{r,T_K} \end{bmatrix},$$

where $T_1,\ldots,T_K$ are the elements of the user set $\mathbf{T}$ in (1), and $\boldsymbol{c}_{i,T_k}=[c_{i,j,T_k}]_{j=1}^{\tilde{F}}$ is a single-column array of length $\tilde{F}$, with j-th entry

$$c_{i,j,T_k} = \begin{cases} *, & \text{if } \tilde{c}_{j,\delta(T_k)} = *, \\ \tilde{c}_{j,\delta(T_k)} + (T_k^{-1}[i]-1)\tilde{S}, & \text{if } \tilde{c}_{j,\delta(T_k)} \neq *. \end{cases}$$

Theorem 2. *Let h,r be positive integers so that $r|h$, and $\tilde{K}=\binom{h-1}{r-1}$. Applying Transformation 1 with parameters (h,r) to a $(\tilde{K},\tilde{F},\tilde{Z},\tilde{S})$ PDA yields a (K,F,Z,S) C-PDA, where*

$$K=\binom{h}{r},\quad F=r\tilde{F},\quad Z=r\tilde{Z},\quad \text{and}\quad S=h\tilde{S}.$$

With the resulting C-PDA, subpacketization level $F=r\tilde{F}$ is sufficient to achieve the rate $R=\frac{S}{Fh}$.

Proof: Array $\boldsymbol{C}$ satisfies C1, C2, and C3 and is thus a PDA. It also satisfies the condition in Definition 3, because $c_{i,j,T_k}=c_{i',j',T_{k'}}=s\in[S]$ implies that $T_k^{-1}[i]=T_{k'}^{-1}[i']$, and thus the labels of all columns containing a given symbol s must have non-empty intersection. The statement on rate and subpacketization follows by Theorem 1 and the discussion before it. ∎

The coding scheme for resolvable combination networks in [7] can be represented in form of a C-PDA, and this C-PDA can be obtained by applying Transformation 1 to the PDA of the Maddah-Ali & Niesen scheme. Theorem 2 thus allows to recover the following result from [7].

Corollary 1. *For a (h,r)-combination network where $r|h$, when $M\in\{0,\frac{Nh}{Kr},\frac{2Nh}{Kr},\cdots,N\}$, there exists a caching scheme that requires rate $R_{\text{TR}}\triangleq\frac{K(1-M/N)}{h(1+KMr/(Nh))}$ and has subpacketization level $F_{\text{TR}}\triangleq r\binom{Kr/h}{KMr/(Nh)}$.*

We apply Transformation 1 to the reduced versions (so as to have the right number of columns) of the low-subpacketization PDAs in Lemmas 1 and 2. This yields the first low-subpacketization C-PDAs and caching schemes for resolvable combination networks.

Theorem 3 (C-PDA construction from Lemma 1). *For any (h,r)-combination network with $r|h$ and cache sizes $M\in\{\frac{1}{q}\cdot N : q\in\mathbb{N}^+, q\geq 2\}$, the following upper bound is achieved by a scheme with subpacketization level $F_{\text{LSub1}}\triangleq r\left(\frac{N}{M}\right)^{\lceil\frac{KMr}{Nh}\rceil-1}$:*

$$R^\star(M)\leq R_{\text{LSub1}}\triangleq\frac{1}{r}\cdot\left(\frac{N}{M}-1\right).$$

(Here, subscript "LSub" stands for "low-subpacketization".)

Proof: By Lemma 1, there exists a PDA with $\lceil\frac{\tilde{K}}{q}\rceil q$ columns. Delete any $\lceil\frac{\tilde{K}}{q}\rceil q-\tilde{K}$ of the columns. Since each ordinary symbol occurs in $\lceil\frac{\tilde{K}}{q}\rceil$ distinct columns, some ordinary symbols can be completely deleted whenever $\lceil\frac{\tilde{K}}{q}\rceil q-\tilde{K}\geq\lceil\frac{\tilde{K}}{q}\rceil$. In this case, the reduced PDA has rate smaller than $\frac{N}{M}-1$. The theorem is concluded by Theorems 1 and 2. ∎

Theorem 4 (C-PDA construction from Lemma 2). *For any (h,r)-combination network with $r|h$ and cache sizes $M\in\{\frac{q-1}{q}\cdot N : q\in\mathbb{N}^+, q\geq 2\}$, the following upper bound is achieved by a scheme with subpacketization level $F_{\text{LSub2}}\triangleq\frac{rM}{N-M}\cdot\left(\frac{N}{N-M}\right)^{\lceil\frac{Kr}{h}\left(1-\frac{M}{N}\right)\rceil-1}$:*

$$R^\star(M)\leq R_{\text{LSub2}}\triangleq\frac{1}{r}\cdot\left(\frac{N}{M}-1\right).$$

Proof: Similarly to the proof of Theorem 3, except that deleting $\lceil\frac{\tilde{K}}{q}\rceil q-\tilde{K}$ columns does not delete any of the ordinary symbols, as each of them occurs $\lceil\frac{\tilde{K}}{q}\rceil(q-1)$ times. ∎

For fair comparison, we compare the new schemes with the scheme in [7] (Corollary 1) when $K\leq N$ for the same memory size. We start with a comparison of the required rates. If $M=\frac{N}{q}$ for some integer $q\geq 2$, then $\frac{KMr}{KMr+Nh}\leq\frac{R_{\text{TR}}}{R_{\text{LSub1}}}\leq 1$. Similarly, if $M=\frac{(q-1)N}{q}$ for some integer $q\geq 2$, then $\frac{KMr}{KMr+Nh}\leq\frac{R_{\text{TR}}}{R_{\text{LSub2}}}\leq 1$. As a consequence, if $M=\frac{N}{q}$ or $M=\frac{(q-1)N}{q}$ for some integer $q\geq 2$, then

$$\lim_{K\to\infty}\frac{R_{\text{TR}}}{R_{\text{LSub1}}}=1\qquad\text{or}\qquad\lim_{K\to\infty}\frac{R_{\text{TR}}}{R_{\text{LSub2}}}=1.$$

On the other hand, for large values of $K\gg 1$, by Corollary 1 and [14, Lemma 4], the subpacketization levels

of the schemes satisfy

$$F_{\text{TR}} \sim \sqrt{\frac{N^2 hr}{2\pi KM(N-M)}} \cdot e^{\frac{Kr}{h}\left(\frac{M}{N}\ln\frac{N}{M}+(1-\frac{M}{N})\ln\frac{N}{N-M}\right)},$$

and

$$F_{\text{LSub1}} \leq re^{\frac{Kr}{h}\cdot\frac{M}{N}\ln\frac{N}{M}},$$
$$F_{\text{LSub2}} \leq \frac{rM}{N-M}e^{\frac{Kr}{h}\cdot(1-\frac{M}{N})\ln\frac{N}{N-M}}.$$

As a consequence, if $M = \frac{N}{q}$ or $M = \frac{(q-1)N}{q}$ for some integer $q \geq 2$, then

$$\varliminf_{K\to\infty} \frac{F_{\text{TR}}}{F_{\text{LSub1}}} = \infty \qquad \text{or} \qquad \varliminf_{K\to\infty} \frac{F_{\text{TR}}}{F_{\text{LSub2}}} = \infty.$$

V. Achieving the Cutset Bound with Low Subpacketization Level

Throughout this section, r, h denote positive integers with $r \leq h$. But r does not necessarily divide h.

Let $S_1, \ldots, S_{\binom{h}{r-1}}$ denote all the subsets of $[h]$ of size $r-1$. Define $\boldsymbol{B}$ as the $\binom{h}{r-1}$-by-$\binom{h}{r}$ dimensional array with element $b_{j,T}$ in row $j \in \{1, \ldots \binom{h}{r-1}\}$ and column $T \in \mathbf{T}$, where

$$b_{j,T} = \begin{cases} *, & \text{if } S_j \not\subset T, \\ T \backslash S_j, & \text{if } S_j \subset T. \end{cases} \tag{3}$$

Notice that the set of arrays $\mathbf{B}$ forms a subset of the PDAs in [15]. They can be proved to be C-PDAs.

Example 3. For $h = 4$ and $r = 2$, the C-PDA $\boldsymbol{B}$ is:

$\{1,2\}$	$\{1,3\}$	$\{1,4\}$	$\{2,3\}$	$\{2,4\}$	$\{3,4\}$
2	3	4	$*$	$*$	$*$
1	$*$	$*$	3	4	$*$
$*$	1	$*$	2	$*$	4
$*$	$*$	1	$*$	2	3

The caching scheme corresponding to the C-PDA $\mathbf{B}$, allows to determine the optimal rate $R^\star(M)$ for large cache sizes M.

Theorem 5. *For an (h, r)-combination network:*

$$R^\star(M) = \frac{1}{r}\left(1 - \frac{M}{N}\right), \quad M \in \left[N\frac{K-h+r-1}{K},\ N\right].$$

This can be achieved with subpacketization level $F = \binom{h}{r-1}$ when $M = N\frac{K-h+r-1}{K}$.

Proof: The converse follows from the cutset lower bound in [6]. For $M = N\left(1 - \frac{h-r+1}{K}\right)$, the upper bound follows by Theorem 1 and the caching scheme corresponding to the C-PDA $\mathbf{B}$ in (3). For $M > N\left(1 - \frac{h-r+1}{K}\right)$, the upper bound follows by time/memory sharing arguments. ∎

The optimal rate $R^\star(M)$ is in general not achieved by the uncoded placement scheme in [7] (see Corollary 1). In fact, at the point $M = N \cdot \left(1 - \frac{h-r+1}{K}\right)$, the scheme in [7] requires rate $R_{\text{TR}} = \frac{1}{r}\left(1 - \frac{M}{N}\right) \cdot \frac{Kr}{Kr-(r-1)(h-r)}$, which is strictly larger than $R^\star(M)$ whenever $r \geq 2$. Moreover, it has a subpacketization level $r\binom{\binom{h-1}{r-1}}{r-1}$, which is significantly higher than the one in Theorem 5.

VI. Conclusion

We introduced the C-PDAs (a subclass of PDAs) to characterize caching schemes with uncoded placement for combination networks. We also proposed a method to transform certain PDAs to C-PDAs for resolvable networks. This allowed us to obtain the first low-subpacketization schemes for resolvable combination networks with a rate that is close to the rate of the uncoded placement schemes in [7]. We also proposed C-PDAs for general combination networks. These C-PDAs have low subpacketization level and achieve the cut-set lower bound when the cache memories are sufficiently large.

Acknowledgment

The work of Q. Yan and M. Wigger has been supported by the ERC Grant *CTO Com*.

References

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

[2] S. S. Bidokhti, M. Wigger, and A. Yener, "Gaussian broadcast channels with receiver cache alignment," in Proc. of *ICC*, 2017, May 2017, Paris, France.

[3] K. H. Ngo, S. Yang, and M. Kobayashi, "Scalable content delivery with coded caching in multi-antenna fading channels," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 548–562, Jan. 2018.

[4] S. P. Shariatpanahi, G. Caire, and B. H. Khalaj, "Multi-antenna coded caching," in Proc. *ISIT*, 2017, pp. 2113–2117, Jul. 2017, Aachen, Germany.

[5] J. Zhang, and P. Elia, "Feedback-aided coded caching for the MISO BC with small caches," in Proc. of *ICC*, 2017, May 2017, Paris, France.

[6] M. Ji, M. F. Wong, A. M. Tulino, J. Llorca, G. Caire, M. Effros, and M. Langberg, "On the fundamental limits of caching in combination networks," in Proc. *SPAWC*, 2015, pp. 695–699, Jun. 2015, Stockholm, Sweden.

[7] L. Tang and A. Ramamoorthy, "Coded caching for networks with resolvability property," in Proc. *ISIT*, 2016, pp. 420–424, Jul. 2016, Barcelona, Spain.

[8] K. Wan, M. Ji, P. Piantanida, and D. Tuninetti, "Novel outer bounds and inner bounds with uncoded cache placement for combination networks with end-user-caches", arXiv:1701.06884v5.

[9] K. Wan, D. Tuninetti, P. Piantanida, and M. Ji, "On combination networks with cache-aided relays and users," arXiv:1803.06123.

[10] A. A. Zewail and A. Yener, "Coded caching for combination networks with cache-aided relays," in Proc. *ISIT*, 2017, pp. 2433–2437, Jun. 2017, Aachen, Germany.

[11] A. A. Zewail and A. Yener, "Combination networks with or without secrecy constraints: the impact of caching relays," arXiv:1712.04930.

[12] K. Wan, D. Tuninetti, M. Ji, and P. Piantanida, "A novel asymmetric coded placement in combination networks with end-user caches," arXiv:1802.10481.

[13] K. Shanmugam, M. Ji, A. M. Tulino, J. Llorca, and A. G. Dimakis, "Finite-length analysis of caching-aided coded multicasting," *IEEE Trans. Inf. Theory,* vol. 62, no. 10, pp. 5524–5537, Oct. 2016.

[14] Q. Yan, M. Cheng, X. Tang, and Q. Chen, "Placement delivery array design for centralized coded caching scheme," *IEEE Trans. Inf. Theory,* vol. 63, no. 9, pp. 5821–5833, Sep. 2017.

[15] Q. Yan, X. Tang, Q. Chen, and M. Cheng, "Placement delivery array design through strong edge coloring of bipartite graphs", *IEEE Commun. Lett.*, vol. 22, no. 2, pp. 236–239, Feb. 2018.

[16] L. Tang, and A. Ramamoorthy, "Low subpacketization schemes for coded caching," in Proc. *ISIT*, 2017, pp. 2790–2794, Jul. 2017, Aachen, Germany.

[17] C. Shangguan, Y. Zhang, and G. Ge, "Centralized coded caching schemes: A hypergraph theoretical approach," arXiv:1608.03989.

[18] K. Shanmuguam, A. M. Tulino, and A. G. Dimakis, "Coded caching with linear subpacketization is possible using Ruzsa-Szeméredi graphs," in Proc. *ISIT*, 2017, pp. 1237–1241, Jul. 2017, Aachen, Germany.

[19] Q. Yan, M. Wigger, and S. Yang, "Placement delivery array design for combination networks with edge caching," arXiv:1801.03048.