# Resource Allocation for Multi-Antenna Coded Caching Systems With Dynamic User Behavior

Milad Abolpour, *Student Member, IEEE*, Mohammadjavad Salehi, *Member, IEEE*, and Antti Tölli, *Senior Member, IEEE*

***Abstract*—In practical cache-aided setups, users may enter or depart the network at any time. The shared caching model can mitigate the detrimental impact of such dynamic user behavior by assigning users to a limited set of caching profiles. In this letter, we propose a highly effective data delivery strategy relying on an uneven subpacketization process to improve the finite-SNR performance of dynamic networks. In order to maximize the utilization of available spatial degrees of freedom, the number of serving users per transmission is fixed to the maximal value, resulting in a larger number of transmissions for users assigned to caching profiles with smaller user counts. Consequently, users assigned to different profiles may receive a varying portion of their requested files in each transmission. To facilitate the efficient delivery of unequal-sized data, an iterative beamformer design given a variable rate allocation per user is also devised. Numerical results demonstrate that the performance of the proposed method is always superior to unicasting and other existing coded caching approaches.**

***Index Terms*—Coded caching, shared caching, dynamic networks, uneven subpacketization, multi-antenna communications.**

## I. Introduction

CODED caching (CC) has been recently proposed as a promising technique to support demanding multimedia applications such as video-on-demand (VoD) and wireless extended reality (XR) [1]. CC leverages the cache memory of network users as a communication resource [2], enhancing the achievable rate of single-stream downlink communications by a multiplicative factor proportional to the cumulative cache size of the entire network. This new CC gain is additive with the spatial multiplexing gain of multi-antenna communications [3], [4], a critical component of existing and upcoming standards [5]. The practical realization of CC gains in multi-input single-output (MISO) setups is, however, restricted by the subpacketization process that entails the fragmentation of each file into numerous smaller parts, with the number of fragments growing exponentially with the number of users $K$ [6], [7], [8]. A practical solution to combat this bottleneck is to implement the shared caching model by assigning $K$ users to $P \leq K$ *caching profiles*, such that each user is assigned to a single profile, and users assigned to the same profile store the same content [9], [10], [11].

Notably, shared caching also provides a viable solution to another critical challenge with CC schemes: the effective management of dynamic networks that experience fluctuations in the user population as the users may enter or depart the network at any time. While conventional CC schemes required a priori knowledge of the number of users for the placement design, the works [12], [13], [14] embraced the shared caching approach to address the dynamicity issue by determining the number of caching profiles solely based on the knowledge of the cache size of users. Specifically, the work [14] presented a universal solution, with delivery algorithms supporting any given network parameters in a dynamic MISO-CC setup and closed-form degrees-of-freedom (DoF) expressions that matched the information-theoretic bounds of [10], [11] under the extremes of small and large spatial multiplexing gains (compared to the CC gain).

Although the DoF metric is an appropriate performance indicator in the high signal-to-noise ratio (SNR) regime, it fails to capture the finite-SNR behavior properly [4], [15]. So, we may still be able to improve the finite-SNR performance of dynamic MISO-CC systems with proper beamforming techniques [4]. Resource allocation in dynamic CC networks has been studied in various works such as [10], [11], [12], [13], [14], [16], [17], [18]. However, unlike the scheme proposed in this letter, their performance is either limited to a specific set of network parameters or they are not able to fully utilize the available spatial multiplexing gain for the data delivery. This letter provides a novel design for the delivery phase that benefits from uneven subpacketization and a flexible, optimized beamformer design to boost the finite-SNR performance of dynamic systems compared to the state-of-the-art. The key idea is to aim at serving a fixed (maximal) number of users in each transmission to allow better balancing of the CC and beamforming gains. The previous schemes relied completely on beamforming techniques to design optimized beams that could benefit from the increased interference-free subspace stemming from the DoF loss incurred by the dynamic network structure [13]. However, due to the logarithmic scaling of the achievable rate w.r.t. the SNR, the relative rate gains from optimized beamforming saturate as the number of users within a given transmission interval is decreased. Hence, the compensation for DoF loss was somewhat limited in the dynamic schemes introduced in [12], [13], [14] as they imposed no control over the number of users served per transmission.

Dynamic user behavior prohibits the wireless network from exploiting the maximal available spatial multiplexing gain due to the non-uniform association of users to profiles. In this letter, the aim is to design a transmission strategy that not only fully utilizes the available spatial multiplexing gain in uneven user-to-profile associations but also improves the

Manuscript received 25 March 2024; revised 7 May 2024; accepted 18 May 2024. Date of publication 22 May 2024; date of current version 9 August 2024. This work was supported in part by the Academy of Finland, 6G Flagship Program under Grant 346208; in part by the Cache-Aided mmWave Access for Immersive Digital Environments (CAMAIDE) under Grant 343586; in part by the Finnish-American Research and Innovation Accelerator (FARIA) Program; and in part by the Tauno Tönning Foundation. The associate editor coordinating the review of this article and approving it for publication was J. Wang. *(Corresponding author: Milad Abolpour.)*

The authors are with the Center for Wireless Communications, University of Oulu, 90570 Oulu, Finland (e-mail: milad.abolpour@oulu.fi).

Digital Object Identifier 10.1109/LWC.2024.3404137

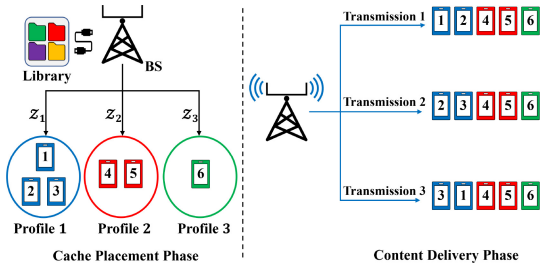



Fig. 1. System model for a dynamic cache-aided MISO network with $\alpha = 2$, $\gamma = \frac{2}{3}$, $\bar{t} = 2$ and $P = 3$. Here, it is assumed that users $\{1, 2, 3\}$, $\{4, 5\}$ and $\{6\}$ are assigned to profiles 1, 2 and 3, respectively, and store the cache content of their corresponding profiles during the placement phase. The delivery phase is comprised of 3 multicast transmissions such that users assigned to profiles 1, 2, and 3 receive 2, 3, and 3 subpackets, respectively.

finite-SNR performance. To this end, we compensate for the network dynamicity using an uneven subpacketization process that allows serving more users in each transmission, close to its maximal value in a comparable shared caching setup with a uniform user-to-profile association [10]. To achieve that, the users assigned to shorter-length caching profiles (i.e., profiles with fewer assigned users) are served by more transmissions but receive smaller chunks of their requested files in each transmission compared to other users. This results in an almost uniform number of users served by each transmission, thus enabling more efficient usage of the available spatial DoF. Of course, as users receive data chunks of unequal sizes, we need multi-rate transmissions. This is addressed in this letter with a fast, iterative beamformer design using a modified version of the design in [14]. Simulation results confirm that the resulting CC scheme for dynamic setups consistently outperforms currently available schemes in the literature.

In this letter, boldface lower-case letters represent vectors. We use $[n]$ to denote the set $\{1, \ldots, n\}$, and $|\mathcal{L}|$ to show the cardinality of $\mathcal{L}$. Also, $\mathcal{L}\|\mathcal{M}$ is the concatenation of $\mathcal{L}$ and $\mathcal{M}$, $(\mathcal{M}\|\mathcal{M})_a$ shows $a$ concatenations of $\mathcal{M}$ with itself,[1] and for $\mathcal{M} \subseteq \mathcal{N}$, $\mathcal{N}\backslash\mathcal{M}$ describes $\mathcal{N} - \mathcal{M}$.

## II. System Model

A dynamic MISO network comprising a varying set of cache-enabled single-antenna users and a base station (BS) with $L$ transmit antennas and the spatial multiplexing gain of $\alpha \leq L$ is considered. The BS has access to a library of $N$ unit-sized files $W^1, W^2, \ldots, W^N$, and each user has a large enough cache memory to store a portion $0 < \gamma < 1$ of the entire library.[2] In this dynamic setup, users can freely enter or depart the network at any moment; hence, the BS does not have any prior information about the number of users to be active during the data delivery. Assuming $\gamma$ is known, the BS selects the parameters $\bar{t}, P \in \mathbb{N}$, ensuring that $\bar{t} = P\gamma$. Upon joining the network, each user $k$ is assigned to a single profile, represented by $\mathsf{p}[k] \in [P]$ (see Fig. 1), and its cache content is updated using a *content placement algorithm*. Let us represent the users assigned to profile $p$ by $\mathcal{U}_p$, i.e., $\mathcal{U}_p = \{k : \mathsf{p}[k] = p\}$. We express the number of users assigned to profile $p \in [P]$ as the length of caching profile $p$ and denote it by $\eta_p$. Without loss of generality, we assume $\eta_1 \geq \eta_2 \geq \cdots \geq \eta_P$. In this letter, we also assume $\alpha \leq \eta_1$. Indeed, for the region $\alpha \leq \eta_1$, adopting the proposed scheme in this letter becomes essential to enhance the finite-SNR performance and enable the system to exploit the available spatial multiplexing gain fully; instead, when $\alpha > \eta_1$, the scheme in [14] can be used to deliver data with the maximal available spatial multiplexing gain.

For the placement phase, we follow a similar approach as in [2], such that each file $W^n$, $n \in [N]$, is split into $\binom{P}{\bar{t}}$ equal-sized mini-files $W^n_{\mathcal{P}}$ as $W^n \to \{W^n_{\mathcal{P}} : \mathcal{P} \subseteq [P], |\mathcal{P}| = \bar{t}\}$. Accordingly, each user $k \in \mathcal{U}_p$ stores the cache content associated with profile $p$, denoted by $\mathcal{Z}_p$ and given by $\mathcal{Z}_p = \{W^n_{\mathcal{P}} : \mathcal{P} \ni p, \mathcal{P} \subseteq [P], |\mathcal{P}| = \bar{t}, \forall n \in [N]\}$.

Due to the dynamic nature of the network, the user population fluctuates over time. During scheduled intervals, active users in the network inform the BS of their desired files from the library. To account for the worst-case scenario, we assume that each file has equal popularity among users, and each user requests a distinct file. Without loss of generality, let us assume that user $k$ requests the file $W^k$. Then, the BS constructs a number of transmission vectors using a *content delivery algorithm* and delivers them to the requesting users, e.g., in consecutive time slots. Every transmission vector $\mathbf{x} \in \mathbb{C}^{L\times 1}$ contains a superposition of a number of precoded codewords, and after its transmission, each target user $k$ receives $y_k = \mathbf{h}_k^{\mathrm{H}}\mathbf{x} + n_k$, where $\mathbf{h}_k \in \mathbb{C}^{L\times 1}$ is the channel vector between the BS and user $k$ and $n_k$ is the zero-mean additive white Gaussian noise of variance $N_0$. In the existing works in the literature [12], [13], [14], to handle non-uniformness in the user-to-profile associations, the BS transmits a different number of subpackets in each transmission, which degrades the rate performance, especially in the finite-SNR regime. In contrast, in this letter, we introduce a novel transmission strategy that fixes the number of transmitted subpackets per transmission to the maximal value. To achieve this, we split each mini-file into an appropriate number of equal-sized subpackets. However, the size of subpackets may vary across different mini-files. The details are provided in the following section.

In this section, we discuss the transmission strategy and resource allocation during the content delivery phase. This phase begins once the active users reveal their required files to the BS. After receiving the users' demands, the BS follows the so-called *flexible subpacketization process* to split each mini-file into several subpackets. The main concept behind this process is to maximize the use of spatial dimensions; thus, the users assigned to profiles with shorter lengths are served in more transmissions, while their rates are adjusted accordingly as their requested files are split into smaller subpackets.

*The Flexible Subpacketization Process:* Let us consider the subset $\mathcal{P} \subseteq [P]$ of profiles with $|\mathcal{P}| = \bar{t}$, such that $\mathcal{P} = \{b_1, b_2, \ldots, b_{\bar{t}}\}$ and $\eta_{b_1} \geq \eta_{b_2} \geq \cdots \geq \eta_{b_{\bar{t}}}$. For each $k \in \mathcal{U}_p$, $p \in [P]\backslash\mathcal{P}$, the mini-file $W^k_{\mathcal{P}}$ is split into $Q^p_{\mathcal{P}}$ subpackets $W^k_{\mathcal{P},q}$, where $Q^p_{\mathcal{P}} = E(\eta_p, \alpha) D_\alpha(\eta_{b_1}, \eta_p) \prod_{i=2}^{\bar{t}} D_\alpha(\eta_{b_i}, \eta_{b_{i-1}})$,

$$E(\eta_p, \alpha) = \begin{cases} \alpha & \eta_p > \alpha \\ 1 & \text{o. w.} \end{cases}, \; D_\alpha(x, y) = \begin{cases} x & x \neq y,\, x > \alpha \\ 1 & \text{o. w.} \end{cases}, \quad (1)$$

and $q \in [Q^p_{\mathcal{P}}]$ increases sequentially to guarantee none of the subpackets are transmitted twice. Here, we note that if $\bar{t} = 1$, it is supposed that $\prod_{i=2}^{\bar{t}} D_\alpha(\eta_{b_i}, \eta_{b_{i-1}}) = 1$, which results in $Q^p_{\mathcal{P}} = E(\eta_p, \alpha) D_\alpha(\eta_{b_1}, \eta_p)$. Next, we present the transmission strategy with flexible subpacketization.

[1] We utilize the generalized multiset definition, in which same elements can be repeated in the sets, e.g., $\{m, m\}$ is not reduced to $\{m\}$.

[2] Clearly, if $\gamma \to 0$, the relative gain of CC w.r.t. to the spatial multiplexing gain vanishes.

### A. Transmission Strategy

Our transmission strategy comprises $\binom{P}{\bar{t}+1}$ rounds, where users from $\bar{t}+1$ profiles are served at each round. We denote the set of profiles selected at round $r$ by $\mathcal{M}(r)$. Every round $r$ also comprises $\rho_r$ individual transmissions, where at each transmission, for every profile $p \in \mathcal{M}(r)$, $\tilde{\eta}_p = \min(\alpha, \eta_p)$ users assigned to profile $p$ are served. Let us assume $\mathcal{M}(r) = \{a_1, a_2, \ldots, a_{\bar{t}+1}\}$ such that $\eta_{a_1} \geq \eta_{a_2} \geq \cdots \geq \eta_{a_{\bar{t}+1}}$. Then, $\rho_r$ is given as

$$\rho_r = \begin{cases} 1 & \eta_{a_1} \leq \alpha \\ \eta_{a_1} \prod_{i=2}^{\bar{t}+1} D_\alpha(\eta_{a_i}, \eta_{a_{i-1}}) & \text{o. w.} \end{cases}. \tag{2}$$

Let us denote the $l$-th transmission of round $r$, $l \in \rho_r$ and $r \in \binom{P}{\bar{t}+1}$, by the *transmission pair* $(r, l)$, and represent the set of users served by the transmission pair $(r, l)$ with $\mathcal{T}_{r,l}$. These user sets are built as follows: for $\mathcal{T}_{r,1}$, we pick the first $\tilde{\eta}_p$ users from every profile $p \in \mathcal{M}(r)$. Then, for $l > 1$, to build $\mathcal{T}_{r,l}$, for every profile $p \in \mathcal{M}(r)$, we circularly shift the user indices that were selected from profile $p$ in transmission pair $(r, l-1)$. In mathematical terms, to build $\mathcal{T}_{r,l}$, for every profile $p \in \mathcal{M}(r)$, we first define the set $\mathcal{V}_p = (\mathcal{V}_{p,1} \| \cdots \| \mathcal{V}_{p,D_\alpha(\eta_p,\alpha)})$, where for $i \in [D_\alpha(\eta_p, \alpha)]$

$$\mathcal{V}_{p,i} = \begin{cases} \bigcup_{j=0}^{\alpha-1} \mathcal{U}_p\big((i+j)\%\eta_p\big) & \eta_p > \alpha \\ \mathcal{U}_p & \text{o. w.} \end{cases}, \tag{3}$$

and $\mathcal{U}_p(i)$ is the $i$-th element of $\mathcal{U}_p$. Here, % sign shows the *elevated* mod operator, for which $a\%\, a = a$ and $(b+a)\%\, a = b\%a$. Indeed, for $i \in [D_\alpha(\eta_p, \alpha)]$, $\mathcal{V}_{p,i}$ makes $i$ circular shifts to $\mathcal{U}_p$, and picks the first $\tilde{\eta}_p$ elements from it. On the other hand, during round $r$, each tuple $\mathcal{V}_{p,i}$ with $p \in \mathcal{M}(r)$ and $i \in [D_\alpha(\eta_p, \alpha)]$ is served in $\frac{\rho_r}{D_\alpha(\eta_p,\alpha)}$ transmissions. Therefore, for each $p \in \mathcal{M}(r)$, we need to expand the set $\mathcal{V}_p$ to $\mathcal{R}_p$, such that $\mathcal{R}_p = (\mathcal{R}_p(1) \| \mathcal{R}_p(2) \| \cdots \| \mathcal{R}_p(\rho_r))$, and for $l \in \rho_r$, $|\mathcal{R}_p(l)| = \tilde{\eta}_p$. Here, $\mathcal{R}_p(l)$ is the $l$-th $\tilde{\eta}_p$-tuple of $\mathcal{R}_p$, indicating the set of users assigned to profile $p \in \mathcal{M}(r)$ and served in the transmission pair $(r, l)$. As a result, we can formulate the set $\mathcal{R}_p$ as follows

$$\mathcal{R}_p = \big(\mathcal{V}_p \| \mathcal{V}_p\big)_{\frac{\rho_r}{D_\alpha(\eta_p,\alpha)}}. \tag{4}$$

Hence, during the transmission pair $(r, l)$, we serve the users assigned to the set $\mathcal{T}_{r,l}$, which is given by:

$$\mathcal{T}_{r,l} = \mathcal{R}_{a_1}(l) \| \mathcal{R}_{a_2}(l) \| \cdots \| \mathcal{R}_{a_{\bar{t}+1}}(l), \tag{5}$$

where $a_i \in \mathcal{M}(r)$ for $i \in [\bar{t}+1]$. Using (5), the BS builds the transmission vector for the transmission pair $(r, l)$ as follows.

$$\mathbf{x}_{r,l} = \sum\nolimits_{i=1}^{\bar{t}+1} \sum\nolimits_{j \in \mathcal{R}_{a_i}(l)} \mathbf{w}_j \, W^j_{\mathcal{M}(r)\setminus\{a_i\},q}, \tag{6}$$

where $\mathbf{w}_j \in \mathbb{C}^{L\times 1}, j \in \mathcal{R}_{a_i}(l)$ are the beamforming vectors that suppress the interference among users in set $\mathcal{R}_{a_i}(l)$ . At the end of this transmission, each user $k$ receives the signal

$$y_k = \sum\nolimits_{i=1}^{\bar{t}+1} \sum\nolimits_{j \in \mathcal{R}_{a_i}(l)} \mathbf{h}_k^{\mathrm{H}} \mathbf{w}_j \, W^j_{\mathcal{M}(r)\setminus\{a_i\},q} + n_k.$$

For $a_i \in \mathcal{M}(r)$ and $a_i \neq \mathsf{p}[k]$, user $k \in \mathcal{T}_{r,l}$ has $W^j_{\mathcal{M}(r)\setminus\{a_i\},q}$ in its cache memory; hence, it can regenerate the interference term $\sum_{i=1, a_i\neq \mathsf{p}[k]}^{\bar{t}+1} \sum_{j\in\mathcal{R}_{a_i}(l)} \mathbf{h}_k^{\mathrm{H}} \mathbf{w}_j \, W^j_{\mathcal{M}(r)\setminus\{a_i\},q}$ and subtract it from $y_k$ to obtain the signal

$$\tilde{y}_k = \sum\nolimits_{j \in \mathcal{R}_{\mathsf{p}[k]}(l)} \mathbf{h}_k^{\mathrm{H}} \mathbf{w}_j \, W^j_{\mathcal{M}(r)\setminus\{\mathsf{p}[k]\},q} + n_k. \tag{7}$$

Here, we note that transmit beamforming vectors $\mathbf{w}_j$ are designed to deliver unequal data lengths. The following example studies the data delivery with flexible subpacketization.

*Example 1:* Let us consider the cache-aided dynamic MISO network depicted in Fig. 1, with $\alpha = 2$, $\gamma = \frac{2}{3}$, $P = 3$, and $\bar{t} = 2$, where users $\mathcal{U}_1 = \{1, 2, 3\}$, $\mathcal{U}_2 = \{4, 5\}$ and $\mathcal{U}_3 = \{6\}$ are assigned to profiles 1, 2 and 3, respectively, while $\eta_1 = 3$, $\eta_2 = 2$ and $\eta_3 = 1$. During the placement phase, each file $W^n$, $n \in [6]$, is split into $\binom{P}{\bar{t}} = 3$ mini-files $W^n_{\mathcal{P}}$ with $\mathcal{P} \in \{\{1,2\},\{1,3\},\{2,3\}\}$ such that user $k \in \mathcal{P}$ stores $W^n_{\mathcal{P}}$ in its cache memory.

For the delivery phase, suppose that users $\{1, \ldots, 6\}$ request the files $W^1, \ldots, W^6$, respectively. In this regard, we have $Q^1_{23} = 2$, $Q^2_{13} = 3$ and $Q^3_{12} = 3$, meaning that each mini-file $W^k_{23}$ is split into 2 subpackets $W^k_{23,q}$ with $k \in \mathcal{U}_1$ and $q \in [2]$, each mini-file $W^k_{13}$ is divided into 3 subpackets $W^k_{13,q}$ with $k \in \mathcal{U}_2$ and $q \in [3]$, and each mini-file $W^k_{12}$ is split into 3 subpackets $W^k_{12,q}$, where $k \in \mathcal{U}_3$ and $q \in [3]$.

During the delivery phase, the transmission strategy consists of $\binom{P}{\bar{t}+1} = 1$ round, and this round is comprised of $\rho_1 = \eta_1 \prod_{i=2}^{3} D_\alpha(\eta_i, \eta_{i-1}) = 3$ transmissions. Here, $\mathcal{M}(1) = \{1, 2, 3\}$, and it is found that $\mathcal{V}_1 = (\{1,2\}\|\{2,3\}\|\{3,1\})$, $\mathcal{V}_2 = (\{4,5\})$ and $\mathcal{V}_3 = (\{6\})$. Then, as per (4), we have $\mathcal{R}_1 = (\{1,2\}\|\{2,3\}\|\{3,1\})$, $\mathcal{R}_2 = (\{4,5\}\|\{4,5\}\|\{4,5\})$ and $\mathcal{R}_3 = (\{6\}\|\{6\}\|\{6\})$.

Now, according to (5), during the transmissions 1, 2 and 3, the users assigned to $\mathcal{T}_{1,1}$, $\mathcal{T}_{1,2}$ and $\mathcal{T}_{1,3}$ are served, respectively, such that (see Fig. 2 for a better clarification):

$$\begin{aligned} \mathcal{T}_{1,1} &= (\mathcal{R}_1(1)\|\mathcal{R}_2(1)\|\mathcal{R}_3(1)) = \{1,2,4,5,6\}, \\ \mathcal{T}_{1,2} &= (\mathcal{R}_1(2)\|\mathcal{R}_2(2)\|\mathcal{R}_3(2)) = \{2,3,4,5,6\}, \\ \mathcal{T}_{1,3} &= (\mathcal{R}_1(3)\|\mathcal{R}_2(3)\|\mathcal{R}_3(3)) = \{1,3,4,5,6\}. \end{aligned}$$

Now, the BS constructs the transmission vectors for the transmissions 1-3. For instance, as per (6), the transmission vector for the first transmission is given by:

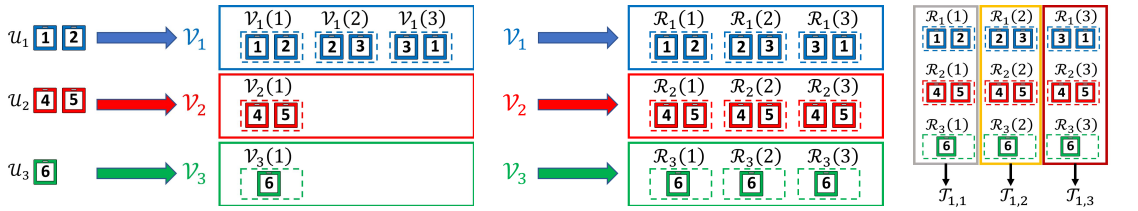
$$\begin{aligned} \mathbf{x}_{1,1} = \sum\nolimits_{i=1}^{3} \sum\nolimits_{j\in\mathcal{R}_i(1)} \mathbf{w}_j \, W^j_{\mathcal{M}(1)\setminus i,q} &= \mathbf{w}_1 \, W^1_{23,1} \\ + \mathbf{w}_2 \, W^2_{23,1} + \mathbf{w}_4 \, W^4_{13,1} + \mathbf{w}_5 \, W^5_{13,1} &+ \mathbf{w}_6 \, W^6_{12,1}, \end{aligned}$$

where $\mathbf{w}_1$, for example, represents the beamforming vector of user 1 that is designed to suppress the interference at user 2 according to the fact that $\mathcal{R}_1(1) = \{1, 2\}$. Finally, as per (7), during the transmission pair (1,1), the received signal at user 1, for example, is given by:

$$\begin{aligned} y_1 = \mathbf{h}_1^{\mathrm{H}} (\mathbf{w}_1 \, W^1_{23,1} + \mathbf{w}_2 \, W^2_{23,1} + \mathbf{w}_4 \, W^4_{13,1} \\ + \mathbf{w}_5 \, W^5_{13,1} + \mathbf{w}_6 \, W^6_{12,1}) + n_1. \end{aligned}$$

Here, user 1 has all subpackets $W^4_{13,1}$, $W^5_{13,1}$ and $W^6_{12,1}$ in its cache memory, therefore, it can regenerate the term $\mathbf{h}_1^{\mathrm{H}}(\mathbf{w}_4 \, W^4_{13,1} + \mathbf{w}_5 \, W^5_{13,1} + \mathbf{w}_6 \, W^6_{12,1})$ and subtract it from $y_1$ to observe the superposition signal $\tilde{y}_1 = \mathbf{h}_1^{\mathrm{H}} \mathbf{w}_1 \, W^1_{23,1} + \mathbf{h}_1^{\mathrm{H}} \mathbf{w}_2 \, W^2_{23,1} + n_1$.

*Remark 1:* The subpacketization complexity of the proposed scheme is $\binom{P}{\bar{t}} \max Q^p_{\mathcal{P}}$, where $\mathcal{P} \subset [P]$, $|\mathcal{P}| = \bar{t}$ and $p \in [P]\setminus\mathcal{P}$. We know that $\alpha \leq \eta_1$ and $\eta_1 \geq \cdots \geq \eta_P$. Hence, as per (1), the subpacketization complexity is upper bounded by $\binom{P}{\bar{t}}\alpha \prod_{i=1}^{\bar{t}} \eta_i$. As mentioned earlier, the existing schemes [10] and [14] with the subpacketization complexity of $\binom{P}{\bar{t}}\alpha$ are also able to handle the data delivery in dynamic networks with $\alpha \leq \eta_1$. However, these schemes cannot fully utilize the available spatial multiplexing gain for data delivery. Notably, although our proposed scheme imposes




(a) Mapping $\mathcal{U}_1 - \mathcal{U}_3$ to $\mathcal{V}_1 - \mathcal{V}_3$ (b) Mapping $\mathcal{V}_1 - \mathcal{V}_3$ to $\mathcal{R}_1 - \mathcal{R}_3$ (c) Building $\mathcal{T}_{1,1} - \mathcal{T}_{1,3}$

Fig. 2. Graphical representation for the resource allocation in Example 1.

further subpacketization complexity compared to [10], [14], it can fully benefit from available spatial multiplexing gain and improves the finite-SNR performance, as it tunes the size of each transmitted data based on the length of caching profiles.

### B. Transmit Beamforming Design

Here, the aim is to assess the finite-SNR behavior based on the symmetric rate performance. To this end, we design flexible optimized beamformers that compensate for unequal data sizes characterized by the dynamic nature of the network. In this regard, for the transmission pair $(r, l)$ with $r \in [\binom{P}{\bar{t}+1}]$ and $l \in [\rho_r]$, as per (7), the achievable rate of user $k \in \mathcal{T}_{r,l}$ is given by $R_k = \log(1+\Gamma_k)$, where $\Gamma_k = \frac{|\mathbf{h}_k^{\mathrm{H}}\mathbf{w}_k|^2}{\sum_{j\in\mathcal{R}_{\mathsf{p}[k]}(l),j\neq k}|\mathbf{h}_k^{\mathrm{H}}\mathbf{w}_j|^2+N_0}$ is the SINR at user $k$. As stated in Section II, during the transmission pair $(r, l)$, user $k \in \mathcal{T}_{r,l}$ receives a $1/\theta_k$ portion of its requested file, where $\theta_k = Q^{\mathsf{p}[k]}_{\mathcal{M}(r)\setminus\{\mathsf{p}[k]\}}\binom{P}{\bar{t}}$. Accordingly, to deliver a subpacket to user $k$ during this transmission, the delivery time $T_k$ is obtained as $T_k = \frac{1}{\theta_k R_k}$. Now, for a given fixed transmit power $P_{\mathrm{T}}$, we solve the following optimization problem to minimize the delivery time or, equivalently, maximize the achievable rate for the worst user served in the transmission pair $(r, l)$ as

$$\max_{\mathbf{w}_k\,:\,k\in\mathcal{T}_{r,l}} \min \ \theta_k R_k, \qquad \text{s.t.} \quad \sum\nolimits_{k\in\mathcal{T}_{r,l}} \|\mathbf{w}_k\|^2 \leq P_{\mathrm{T}}, \tag{8}$$

which is a weighted max-min optimization problem efficiently solved in various studies, such as [6], [14], [19]. For instance, by using a Lagrangian-duality-based technique [14, eq. (22)], we can first write $\mathbf{w}_k = \sqrt{\tilde{p}_k}\tilde{\mathbf{w}}_k$ with $\|\tilde{\mathbf{w}}_k\| = 1$, and then, follow a similar way as in [14, eq. (23)] to get the closed-form expression for the normalized beamformer $\tilde{\mathbf{w}}_k$ as $\tilde{\mathbf{w}}_k = \tilde{}_k^{-1}\mathbf{h}_k / \|\tilde{}_k^{-1}\mathbf{h}_k\|$, where $\tilde{}_k = \mathbf{I}_L + \sum_{j\in\mathcal{T}_{r,l}} \lambda_j \mathbf{h}_j\mathbf{h}_j^{\mathrm{H}}$, and $\lambda_j$ is the dual variable corresponding to beamformer $\tilde{\mathbf{w}}_j$ found by fixed-point iteration [20]. We note that the computation complexity for designing the beamforming vector $\mathbf{w}_k$ is $\mathcal{O}(L^3 + |\mathcal{T}_{r,l}|^3)$. Accordingly, the beamforming power $\tilde{p}_k$ is computed by using the Lagrangian duality, as outlined in [6, eq. (26)]. Now, for each transmission pair $(r, l)$ with $r \in [\binom{P}{\bar{t}+1}]$ and $l \in [\rho_r]$, we compute the transmission rate $R_{r,l} = \min_{k\in\mathcal{T}_{r,l}} \log(1+\Gamma_k)$. Finally, in order to assess the system performance, we can use the *symmetric rate* inversely proportional to the total delivery time, where the transmission time for each transmission pair $(r, l)$ is minimized by maximizing $R_{r,l}$ in the optimization problem (8). Accordingly, the symmetric rate, defined as the total number of bits per second delivered to each user in the delivery phase, is obtained as $R_{\mathrm{sym}} = (\sum_{r=1}^{\binom{P}{\bar{t}+1}} \sum_{l=1}^{\rho_r} \frac{1}{R_{r,l}})^{-1}$. It is worth mentioning that any beamforming design, linear and non-linear, can be applied to the proposed transmission strategy to maximize the achievable symmetric rate. For instance, the linear beamforming design in problem (8) can be replaced by the non-linear beamformers based on dirty paper coding (DPC), where the optimal precoders may be solved via uplink-downlink duality.

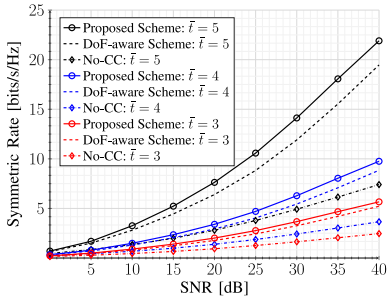



Fig. 3. Symmetric rate versus SNR for different $\bar{t}$ values, where $P = 6$, $\sigma = 2.29$ and $L = \alpha = 4$.

## III. Numerical Results

We consider a dynamic network with $K = 32$ cache-enabled users, where the users are randomly assigned to $P$ caching profiles. Note that, for a given $\gamma$ value, considering any value for the number of files $N \geq K$ does not alter the simulation results, as both the placement and delivery phases do not require the exact knowledge of $N$. To measure the non-uniformness of the user-to-profile association, we use the standard deviation parameter, defined as $\sigma^2 = \frac{1}{P}\sum_{p=1}^{P}(\eta_p - \eta_{\mathrm{avg}})^2$ with $\eta_{\mathrm{avg}} = \frac{K}{P}$. Depending on the scenario, simulations are done for three schemes: 1) the proposed scheme in this letter with uneven subpacketization and optimized beamformers, 2) the scheme in [14], referred to as *DoF-aware*, where the optimized beamformer design of [6] is also applied, and 3) a baseline scheme, dubbed as *No-CC*, where no coded caching technique is applied (the cache contents are used only for a local gain) but the optimized beamformer design of [6] is still used to improve the rate. The No-CC scheme consists of $K$ transmissions such that users $\{1, 2, \ldots, \alpha\}$ are served during the first transmission, and for each subsequent transmission, the served user indices are a circular shift of the user indices of the previous transmission.

Fig. 3 compares the performance of various schemes for different values of $\bar{t}$ and SNR when $P = 6$, $\sigma = 2.29$, and $L = \alpha = 4$. Accordingly, in Fig. 4, the impacts of spatial multiplexing gain $\alpha$ on the symmetric rate performance of our proposed scheme and the DoF-aware approach are depicted in a dynamic setup with $L = \alpha$, $P = 3$, $\sigma = 3.77$ and $\bar{t} = 2$. As observed, the No-CC scheme has the worst performance among the three schemes as it lacks any coded caching gain. Furthermore, it is demonstrated that the proposed scheme outperforms the scheme of [14] over the whole SNR range and for every value of $\bar{t}$ and $\alpha$. This clarifies the improvements resulting from the new design of the delivery

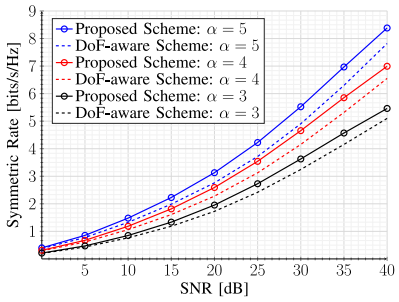



Fig. 4. Symmetric rate versus SNR for different $\alpha$ values with $L = \alpha$, $P = 3$, $\sigma = 3.77$ and $\bar{t} = 2$.

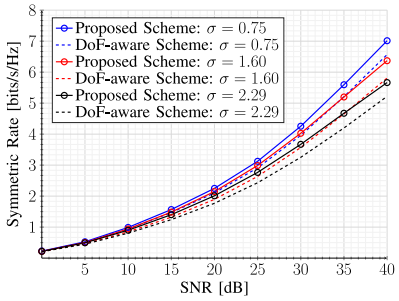



Fig. 5. Symmetric rate versus SNR for different $\sigma$ values with $P = 6$, $\bar{t} = 3$ and $L = \alpha = 4$.

scheme that compensates for the non-uniformness in user-to-profile association with uneven subpacketization.

For better clarification, let us take a closer look at the performance comparison of the proposed and DoF-aware schemes in Figs. 3 and 4. Recall that to handle non-uniform user-to-profile assignments, both schemes select $\bar{t}+1$ profiles during each transmission. Let $p$ be a selected profile, and consider the transmission vectors that select profile $p$. While the uneven subpacketization policy of our scheme enables serving $\min(\alpha, \eta_p)$ users assigned to profile $p$ with *every* such transmission vector, with the DoF-aware scheme, some vectors may not include any users assigned to profile $p$ due to the even subpacketization requirement. So, the average number of users served with the transmission vectors is larger in our scheme, and the gap in the number of users served by the two schemes is proportional with $\min(\alpha, \eta_p)$. This clarifies the performance comparisons: in Fig. 3, by increasing $\bar{t}$, the number of profiles served per transmission is increased linearly, causing the performance gap to grow also visibly larger. However, in Fig. 4, as we increase $\alpha$, the performance gap grows only slightly larger as the increase in the average number of users served per transmission is proportional with $\min(\alpha, \eta_p)$ which is capped by $\eta_p$.

Finally, Fig. 5 illustrates the impacts of non-uniformness in the user-to-profile association on the performance of the proposed scheme and the scheme of [14]. As observed, while the performance of both schemes deteriorates as the user-to-profile association becomes more non-uniform, our scheme still outperforms the DoF-aware scheme of [14] over the entire SNR range, and their performance gap is almost independent of how non-uniform the association is.

## IV. Conclusion

In this letter, we presented a novel data delivery algorithm to boost the finite-SNR performance of cache-aided content distribution in dynamic MISO networks. Our solution optimized the utilization of spatial DoF by fixing the number of users served per transmission at a maximal value derived from comparable shared caching setups with a uniform user-to-profile association. This led to variable-sized file portions being delivered to different users in each transmission. Accordingly, an iterative beamformer design with multi-rate transmissions was employed to facilitate the delivery of unequal-sized data. It was shown through numerical simulations that our proposed scheme consistently outperforms state-of-the-art solutions for data delivery in dynamic MISO setups.

## References

[1] M. Salehi, K. Hooli, J. Hulkkonen, and A. Tölli, "Enhancing next-generation extended reality applications with coded caching," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 1371–1382, 2023.

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

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

[4] A. Tolli, S. P. Shariatpanahi, J. Kaleva, and B. H. Khalaj, "Multi-antenna interference management for coded caching," *IEEE Trans. Wireless Commun.*, vol. 19, no. 3, pp. 2091–2106, Mar. 2020.

[5] N. Rajatheva et al., "White paper on broadband connectivity in 6G," 2020, *arXiv:2004.14247*.

[6] M. J. Salehi, E. Parrinello, S. P. Shariatpanahi, P. Elia, and A. Tölli, "Low-complexity high-performance cyclic caching for large MISO systems," *IEEE Trans. Wireless Commun.*, vol. 21, no. 5, pp. 3263–3278, May 2022.

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

[8] H. H. S. Chittoor, P. Krishnan, K. V. Sushena Sree, and M. V. N. Bhavana, "Subexponential and linear subpacketization coded caching via projective geometry," *IEEE Trans. Inf. Theory*, vol. 67, no. 9, pp. 6193–6222, Sep. 2021.

[9] S. Jin, Y. Cui, H. Liu, and G. Caire, "A new order-optimal decentralized coded caching scheme with good performance in the finite file size regime," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5297–5310, Aug. 2019.

[10] E. Parrinello, A. Ünsal, and P. Elia, "Fundamental limits of coded caching with multiple antennas, shared caches and uncoded prefetching," *IEEE Trans. Inf. Theory*, vol. 66, no. 4, pp. 2252–2268, Apr. 2020.

[11] E. Parrinello, P. Elia, and E. Lampiris, "Extending the optimality range of multi-antenna coded caching with shared caches," in *Proc. IEEE ISIT*, Los Angeles, CA, USA, 2020, pp. 1675–1680.

[12] M. Abolpour, M. J. Salehi, and A. Tölli, "Coded caching and spatial multiplexing gain trade-off in dynamic MISO networks," in *Proc. IEEE SPAWC*, Oulu, Finland, 2022, pp. 1–5.

[13] M. Salehi, E. Parrinello, H. B. Mahmoodi, and A. Tölli, "Low-subpacketization multi-antenna coded caching for dynamic networks," in *Proc. IEEE EuCNC/6G Summit*, Grenoble, France, 2022, pp. 112–117.

[14] M. Abolpour, M. Salehi, and A. Tölli, "Cache-aided communications in MISO networks with dynamic user behavior," *IEEE Trans. Wireless Commun.*, early access, May 9, 2024, doi: 10.1109/TWC.2024.3395368.

[15] M. Salehi and A. Tölli, "Multi-antenna coded caching at finite-SNR: Breaking down the gain structure," in *Proc. IEEE Asilomar Conf. Signals Syst. Comput.*, Pacific Grove, CA, USA, 2022, pp. 703–708.

[16] Y. Ma and D. Tuninetti, "On coded caching systems with offline users," in *Proc. IEEE ISIT*, Espoo, Finland, 2022, pp. 1133–1138.

[17] E. Lampiris and P. Elia, "Adding transmitters dramatically boosts coded-caching gains for finite file sizes," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 6, pp. 1176–1188, Jun. 2018.

[18] H. Zhao, A. Bazco-Nogueras, and P. Elia, "Vector coded caching multiplicatively increases the throughput of realistic downlink systems," *IEEE Trans. Wireless Commun.*, vol. 22, no. 4, pp. 2683–2698, Apr. 2023.

[19] H. B. Mahmoodi, M. Salehi, and A. Tölli, "Non-symmetric multi-antenna coded caching for location-dependent content delivery," in *Proc. IEEE ICC*, Seoul, South Korea, 2022, pp. 5165–5170.

[20] A. Wiesel, Y. C. Eldar, and S. Shamai, "Linear precoding via conic optimization for fixed MIMO receivers," *IEEE Trans. Signal Process.*, vol. 54, no. 1, pp. 161–176, Jan. 2006.