Cache-Aided Communications in MISO Networks with Dynamic User Behavior: A Universal Solution

Milad Abolpour, MohammadJavad Salehi, and Antti Tölli

Centre for Wireless Communications, University of Oulu, 90570 Oulu, Finland
E-mail: {firstname.lastname}@oulu.fi

***Abstract*—A practical barrier to the implementation of cache-aided networks is dynamic and unpredictable user behavior. In dynamic setups, users can freely depart and enter the network at any moment. The shared caching concept has the potential to handle this issue by assigning K users to P caching profiles, where all η_p users assigned to profile p store the same cache content defined by that profile. The existing schemes, however, cannot be applied in general and are not dynamic in the true sense as they put constraints on the transmitter-side spatial multiplexing gain α. Specifically, they work only if $\alpha \leq \min_p \eta_p$ or $\alpha \geq \hat{\eta}$, where in the latter case, γ is the normalized cache size of each user, $\hat{\eta}$ is an arbitrary parameter satisfying $1 \leq \hat{\eta} \leq \max_p \eta_p$, and the extra condition of $\alpha \geq K\gamma$ should also be met. In this work, we propose a universal caching scheme based on the same shared-cache model that can be applied to any dynamic setup, extending the working region of existing schemes to networks with $\min_p \eta_p \leq \alpha \leq \hat{\eta}$ and removing any other constraints of existing schemes. We also derive the closed-form expressions for the achievable degrees-of-freedom (DoF) of the proposed scheme and show that it achieves the optimal DoF for uniform user distributions. Notably, it is the first scheme to achieve the optimal DoF of $K\gamma + \alpha$ for networks with uniform user distribution, $\alpha > \hat{\eta}$, and non-integer $\frac{\alpha}{\hat{\eta}}$, without imposing any other constraints. Finally, we use numerical simulations to assess how non-uniform user distribution impacts the DoF performance and illustrate that the proposed scheme provides a noticeable improvement over unicasting for uneven distributions.**

***Index Terms*—coded caching; shared caching; dynamic networks; multi-antenna communications**

I. Introduction

The increasing volume and diversity of multimedia content require wireless networks to be enhanced to serve users at higher data rates and with lower latency [1], while network providers must further develop infrastructure in anticipation of evolving applications such as wireless immersive viewing [2], [3]. To facilitate the efficient delivery of such multimedia content, coded caching (CC) has been proposed to increase the data rates by leveraging the cache memory across the network as a communication resource [4]. Accordingly, incorporating CC into a single-stream downlink network boosts the achievable rate by a multiplicative factor proportional to the cumulative cache size in the entire network via multicasting carefully designed codewords to different user groups. In light of the significance of multi-antenna connectivity in deploying next-generation networks [1], various works have studied the performance of cache-aided multi-input single-output (MISO) configurations [5]–[10]. For instance, [8] and [11] discussed the design of optimized beamformers in finite signal-to-noise-ratio (SNR), and [2] explored the capability of CC to cope with location-dependent file request applications.

In practice, however, practically achievable CC gains are constrained by the subpacketization process [12]–[14]. That is, in a network with K users, each file should be split into many smaller parts, the number of which grows exponentially with K. A promising way to overcome this impediment is to use the shared caching concept, where there exist $P \leq K$ *caching profiles*, and η_p users are assigned to profile $p \in \{1, \cdots, P\}$. Even though with this concept, multiple users with a cache ratio of γ could be assigned to the same profile and cache exactly the same data, in [15], it is shown that in MISO setups with $\alpha \geq \frac{K}{P}$, the scaling factor in the degrees-of-freedom (DoF) could be the same as the case of dedicated users' caches, i.e., $K\gamma+\alpha$, where α is the spatial multiplexing gain. However, for a shared-cache MISO setup with $\alpha \leq \frac{K}{P}$, the optimal DoF is $\alpha(1 + P\gamma)$ [16].

Interestingly, shared caching can also address another critical issue with coded caching schemes: handling networks with a dynamic population of users departing and entering the network at any time. The problem with conventional CC schemes is that they require the placement phase to be designed based on the number of users known a priori. By contrast, this problem is alleviated with the shared-cache model since the cache placement phase is built upon knowledge of the number of profiles P, and not the number of users K. Accordingly, in some cache-aided scenarios, such as extended reality applications [3], the server is aware of the cache ratio of users rather than the number of existing users. In this sense, the authors in [17] and [18], apply the shared caching idea to address the dynamicity issue. In this method, the server only needs to know the cache ratio of users to determine the number of caching profiles and design the content placement phase. Although shared caching is crucial for managing dynamic conditions, the existing models are not dynamic in the true sense and only support two regions: *1)* $\alpha \leq \min_p \eta_p$ [16], and *2)* $\alpha \geq \hat{\eta}$ with $\alpha \geq K\gamma$ and arbitrary $\hat{\eta}$ satisfying

This research has been supported by the Academy of Finland, 6G Flagship program under Grant 346208, 343586 (CAMAIDE), and by the Finnish-American Research and Innovation Accelerator (FARIA) program.

$1 \leq \hat{\eta} \leq \max_p \eta_p$ [17], [18]. Therefore, a universal shared-cache setup that supports any user-to-profile association is not yet available.

In this work, we design a universal cache-assisted MISO system capable of handling any instantaneous user distribution among caching profiles. Our system operation is comprised of two phases: *i) Content placement phase*, and *ii) content delivery phase*. In the placement phase, the server determines the number of caching profiles according to the cache ratio γ. Then, upon connecting to the network, each user is assigned to a single profile and stores the cache content of that profile. During the content delivery phase, the server employs a clever combination of multicast and unicast transmissions to maximize the DoF. In this paper, we obtain closed-form expressions for the DoF, revealing the DoF loss caused by non-uniformness in users' distribution. Particularly, for the uniform user associations, it is shown that our proposed scheme achieves the optimal DoF not only in the regions covered in the literature but also in the region $\alpha \geq \hat{\eta}$ with non-integer $\alpha/\hat{\eta}$. Notably, our proposed scheme supports any user distribution, including the regions omitted in the existing literature [15]–[17] such as networks with: *i)* uneven user association with $\alpha > \hat{\eta}$ and non-integer $\frac{\alpha}{\hat{\eta}}$ unlike [15], *ii)* $\min_p \eta_p \leq \alpha \leq \hat{\eta}$ unlike [16], and *iii)* $\alpha \geq \hat{\eta}$ with $\alpha < K\gamma$ unlike [17].

In this paper, bold lower-case and calligraphic letters show vectors and sets, respectively. $[a:b]$ shows the set $\{a, \cdots, b\}$, $[a] = \{1, \cdots, a\}$, $|\mathcal{A}|$ is the cardinality of $\mathcal{A}$, and for $\Lambda \subseteq \mathcal{A}$, $\mathcal{A}_{\backslash \Lambda}$ represents $\mathcal{A} - \Lambda$. $(\mathcal{A}\|\mathcal{A})_x$ denotes x concatenations of $\mathcal{A}$ with itself, and $\mathcal{A}\|\mathcal{B}$ is the concatenation of $\mathcal{A}$ and $\mathcal{B}$.

II. System Model

In this paper, we focus on a dynamic MISO network, where a base station (BS) equipped with L transmit antennas and the spatial multiplexing gain of $\alpha \leq L$ serves several cache-enabled single-antenna users. The BS has access to a library $\mathcal{F}$ with N equal-sized files, and each user is equipped with a large enough memory to store a portion $0 < \gamma < 1$ of the entire library. We suppose $\gamma = \frac{\bar{t}}{P}$, where $\bar{t}$ and P are natural numbers and $\gcd(\bar{t}, P) = 1$. In this dynamic setup, users can move, enter and depart the network at any time. Accordingly, the BS does not have any prior knowledge about the number of available users during the transmission. When a user k enters the network, it is assigned to a profile represented by $\mathsf{p}[k] \in [P]$, and the content of its cache is updated based on a *content placement algorithm*.

In the placement phase, by following the same way as in [4], each file $W^n \in \mathcal{F}$, $n \in [N]$, is split into $\binom{P}{\bar{t}}$ equal-sized mini-files $W^n_{\mathcal{P}}$ such that $W^n \rightarrow \{W^n_{\mathcal{P}} : \mathcal{P} \subseteq [P], |\mathcal{P}| = \bar{t}\}$. The cache content associated with profile $p \in [P]$, represented by $\mathcal{Z}_p$, includes a portion γ of each file W^n as

$$\mathcal{Z}_p = \{W^n_{\mathcal{P}} : \mathcal{P} \ni p, \mathcal{P} \subseteq [P], |\mathcal{P}| = \bar{t}, \forall n \in [N]\}.$$

Then, defining $\mathcal{U}_p$ as the set of users assigned to profile p, i.e., $\mathcal{U}_p = \{k : \mathsf{p}[k] = p\}$, each user $k \in \mathcal{U}_p$ stores the cache content $\mathcal{Z}_p$ during the placement phase.

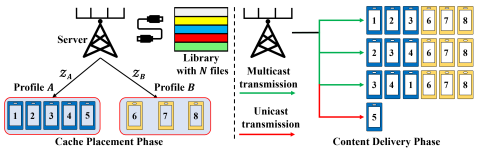


Fig. 1. System model for a dynamic coded caching setup, where $P = 2$, $\gamma = \frac{1}{2}$, $\bar{t} = 1$, $\alpha = 4$, $\hat{\eta} = 4$, $Q = 2$ and $\beta = 3$. During the placement phase, each user assigned to profiles A and B stores the cache content associated with those profiles. For the delivery phase, user 5 is served via unicasting and other users are served via 3 multicast transmissions.

The dynamic nature of the network causes a fluctuating user population throughout time. During regular intervals, the network's demanding users reveal their required files from the library $\mathcal{F}$ to the BS. Then, using a *content delivery algorithm*, the BS constructs and transmits a set of codewords, enabling users to retrieve their requested files. In this paper, we focus on the content delivery procedure over a specific time interval, where it is assumed that the number of present users during the BS's transmission is K. In line with the general approach in the literature, we utilize the total DoF as the metric of interest, representing the average number of concurrent users served in parallel across all transmit intervals. The main contribution of this paper is to design delivery algorithms that provide a maximum combination of global caching and spatial multiplexing gains under the proposed dynamic conditions. As part of the delivery process, we discuss the transmission strategies to reduce the DoF loss caused by non-uniformness in user-to-profile association in the following section.

III. Resource Allocation and Data Delivery

In this section, we discuss the resource allocation and transmission strategies during the content delivery phase. This phase commences once the set of active users reveals their requested files and comprises two consecutive steps: *1) Coded caching (CC) data delivery*; and *2) Unicast (UC) data delivery*.

Let us define the number of users assigned to profile p as the length of profile p denoted by η_p, where without loss of generality, it is assumed that $\eta_1 \geq \eta_2 \geq \cdots \geq \eta_P$. By choosing a *delivery parameter* $\hat{\eta} \leq \max_p \eta_p$[1], the BS builds and transmits a set of codewords to serve at most $\hat{\eta}$ users assigned to each profile with a novel CC-based approach. In this regard, for every profile p

- if $\hat{\eta} < \eta_p$, we exclude $\eta_p - \hat{\eta}$ users, and exempt BS to serve these users during the CC delivery step. Accordingly, the excluded users are served in the UC delivery step.
- if $\hat{\eta} \geq \eta_p$, all η_p users are served via the CC delivery step.

Now, let us suppose that the set of users assigned to profile $p \in [P]$ and served during the CC delivery step is denoted by $\mathcal{V}_p$ such that $|\mathcal{V}_p| = \delta_p$, $\mathcal{V}_p = \{v_{p,1}, v_{p,2} \cdots, v_{p,\delta_p}\}$, and $v_{p,i} \in \mathcal{U}_p$ for $i \in [\delta_p]$. We note that $\delta_p = \min(\hat{\eta}, \eta_p)$, and clearly, $\delta_1 = \hat{\eta}$ and $\delta_1 \geq \delta_2 \geq \cdots \geq \delta_P$.

In order to build the transmission vectors, as depicted in Fig. 1, the BS selects a parameter Q, which represents

[1]Here, the delivery parameter plays a similar role as the unifying length parameter in [17]. Both delivery and unifying length parameters tune the DoF loss caused by the non-uniformness in the user-to-profile association.

the number of profiles served in each transmission, and a parameter β, which expresses the number of users chosen from each profile to serve in each transmission. The necessary conditions for choosing any arbitrary values for Q and β are defined in Remark 1.

Remark 1: In order to serve Q profiles each with a maximum of β users, the network parameters should satisfy the constraints $\bar{t}+1 \leq Q \leq \bar{t}+\lceil \alpha/\beta \rceil$ and $\beta \leq \min(\alpha, \hat{\eta})$.

Proof: The proof is relegated to [19, Appendix A]. ■

In the following, we present the system operation maximizing the DoF performance separately for two regimes: *i)* $\alpha \leq \hat{\eta}$ and *ii)* $\alpha > \hat{\eta}$. For the CC delivery step, each of these cases operates either with *Strategy A* (cf. Section III-A) or *Strategy B* (cf. Section III-B). The chosen transmission strategy only depends on the parameters α and $\hat{\eta}$, and it is independent of the content placement phase. In other words, the server performs the placement phase only based on parameter γ without considering which transmission strategy the system will use for the CC delivery step.

1) System operation for $\alpha \leq \hat{\eta}$: For the case of $\alpha \leq \hat{\eta}$, we set $\beta = \alpha$, and the only option for Q to maximize the DoF is $Q = \bar{t}+1$. Here, *Strategy A* is utilized to build the transmission vectors. Replacing the constraint $\alpha \leq \hat{\eta}$ with $\alpha \leq \min_p \delta_p$ reduces our system model to [16], while our proposed scheme also works for the scenarios with $\min_p \delta_p < \alpha \leq \hat{\eta}$.

2) System operation for $\alpha > \hat{\eta}$: For this case, we set $\beta = \hat{\eta}$, and define $\hat{\alpha} = \frac{\alpha}{\hat{\eta}}$. If $\hat{\alpha}$ is an integer, then we follow *Strategy A* to build the transmission vectors. For non-integer $\frac{\alpha}{\hat{\eta}}$, the server can serve users via *Strategy A* by setting $\bar{t}+1 \leq Q \leq \bar{t}+\lfloor \alpha/\hat{\eta} \rfloor$, and via *Strategy B* by choosing $Q = \bar{t}+\lceil \alpha/\hat{\eta} \rceil$. In this case, if we assume that users are uniformly distributed among caching profiles, and $\frac{\alpha}{\hat{\eta}}$ is an integer, the system performance is simplified to [15], while our proposed scheme also covers the uneven user-to-profile associations with non-integer $\frac{\alpha}{\hat{\eta}}$.

A. Transmission Strategy A

In this strategy, each mini-file $W^n_{\mathcal{P}}$ is split into $\beta\binom{P-\bar{t}-1}{Q-\bar{t}-1}$ subpackets $W^n_{\mathcal{P},q}$, where $q \in [\beta\binom{P-\bar{t}-1}{Q-\bar{t}-1}]$ increases sequentially after each transmission to ensure none of subpackets is transmitted twice. Next, we follow the so-called *elevation process* to serve Q caching profiles each with at most β users.

Elevation process: In this process, the aim is to characterize the set of users that are served in each transmission. Accordingly, we let $\phi_p = \max(\beta, \delta_p)$, and use the so-called *transmission triple* (r, c, l), where $r \in [P-Q+1]$, $c \in [\phi_r]$ and $l \in \left[\binom{P-r}{Q-1}\right]$. Also, it is assumed that $\eta_1 \geq \eta_2 \geq \cdots \geq \eta_P$, which results in $\phi_1 \geq \phi_2 \geq \cdots \geq \phi_P$. This process creates $\mathcal{T}_{r,c,l}$, the set of users that are served during the transmission triple (r, c, l). In this regard, first, for every $p \in [P]$, we elevate the set $\mathcal{V}_p$ to the set $\mathcal{R}_p$ as follows.

$$\mathcal{R}_p = \mathcal{R}_{p,1} \| \cdots \| \mathcal{R}_{p,\phi_p}, \tag{1}$$

where, for $j \in [\phi_p]$, $\mathcal{R}_{p,j}$ is defined as:

$$\mathcal{R}_{p,j} = \begin{cases} \mathcal{V}_p & \delta_p \leq \beta \\ \{v_{p,l} : l = (i+j-1)\%\delta_p, 1 \leq i \leq \beta\} & \delta_p > \beta \end{cases}.$$

Here, $\%$ sign demonstrates the mod operator, for which $c\%c = c$ and $(d+c)\%c = d\%c$. Furthermore, we use generalized multiset definition where we allow the same elements to be repeated in the sets, e.g., $\{a, a\}$ cannot be reduced to $\{a\}$. Now, for $p \in [P]$, we define $\mathcal{S}_p = \mathcal{S}_{p,1} \| \cdots \| \mathcal{S}_{p,\hat{\eta}}$, while

$$\mathcal{S}_{p,j} = \begin{cases} \mathcal{R}_{p,j} & 1 \leq j \leq \phi_p \\ \varnothing & \phi_p + 1 \leq j \leq \hat{\eta} \end{cases}. \tag{2}$$

Then, for each $r \in [P-Q+1]$, we define the set $\mathcal{M}_r$ as:

$$\mathcal{M}_r = \{\mathcal{F} : \mathcal{F} \subseteq [r+1:P], |\mathcal{F}| = Q-1\}. \tag{3}$$

In the proceeding, we use $\mathcal{M}_r(l)$ to indicate the l-th $(Q-1)$-tuple of $\mathcal{M}_r$. Finally, the set $\mathcal{T}_{r,c,l}$ for the transmission triple (r, c, l) is given by:

$$\mathcal{T}_{r,c,l} = \big\{\mathcal{S}_{r,c} \| \mathcal{S}_{b_1,c} \| \cdots \| \mathcal{S}_{b_{Q-1},c} : b_i \in \mathcal{M}_r(l), \forall i \in [Q-1]\big\}. \tag{4}$$

Generally speaking, for the transmission triple (r, c, l), if $\delta_r = 0$, the BS does not transmit any signal; otherwise, the BS serves users assigned to $\mathcal{T}_{r,c,l} = \mathcal{S}_{r,c} \| \mathcal{S}_{b_1,c} \| \cdots \| \mathcal{S}_{b_{Q-1},c}$, while $\{b_1, \cdots, b_{Q-1}\} = \mathcal{M}_r(l)$. During the transmission triple (r, c, l), by defining $\mathcal{N} = \{r\} \cup \mathcal{M}_r(l)$, the BS constructs the transmission vector as follows.

$$\mathbf{x}_{r,c,l} = \sum_{\Lambda \subseteq \mathcal{N}: |\Lambda| = \bar{t}} \; \sum_{k \in \mathcal{S}_{p,c}: p \in \mathcal{N}_{\setminus \Lambda}} W^k_{\Lambda,q} \mathbf{w}_{\mathcal{G}^k_\Lambda},$$

where $\mathbf{w}_{\mathcal{G}^k_\Lambda} \in \mathbb{C}^{L \times 1}$ is the zero-forcing (ZF) precoder that cancels out the interference of user k at the set $\mathcal{G}^k_\Lambda = \big\{j \in \mathcal{S}_{p,c} : \forall p \in \mathcal{N}_{\setminus \Lambda}, j \neq k\big\}$. In [19, Appendix D], it is proven that all users served with *Strategy A* can decode their requested files at the end of the CC delivery step. Now, assuming $\mathbf{h}_i \in \mathbb{C}^{L \times 1}$ as the channel gain of user i, at the end of the transmission triple (r, c, l), the received signal at user i is given by:

$$y_i = \mathbf{h}_i^{\mathrm{H}} \mathbf{x}_{r,c,l} + n_i,$$

where n_i is the zero-mean additive white Gaussian noise of unit variance. In order to give further insight, an example for data delivery with *Strategy A* is provided in [19, Appendix B].

B. Transmission Strategy B

When $\alpha > \hat{\eta}$ and $\frac{\alpha}{\hat{\eta}}$ is not an integer, we can serve users by setting $\beta = \hat{\eta}$, and $Q = \bar{t} + \lceil \alpha/\hat{\eta} \rceil$.[2] Here, first, we split each mini-file into $(\hat{\eta}\bar{t}+\alpha)\binom{P-\bar{t}-1}{Q-\bar{t}-1}\binom{Q-2}{Q-\bar{t}-2}$ subpackets. Then, we use the elevation process to serve Q profiles each with at most β users.

Elevation process: This process builds the set of users served in each transmission. First, for $r \in [P]$, we define $\mathcal{Y}_r$ as:

$$\mathcal{Y}_r = \begin{cases} \mathcal{V}_r & \delta_r = \hat{\eta} \\ \mathcal{V}_r \| (f^* \| f^*)_{\hat{\eta}-\delta_r} & \text{o.w.} \end{cases}, \tag{5}$$

where f^* denotes the phantom (non-existent) users. Generally speaking, in each transmission of *Strategy B*, we serve the

[2]For $\alpha > \hat{\eta}$ and non-integer $\frac{\alpha}{\hat{\eta}}$, we can still serve users with *Strategy A*. However, the achievable DoF is less than the one with *Strategy B*.

users assigned to Q profiles such that we select at most $\hat{\eta}$ users from $Q-1$ profiles, and pick $\theta = \alpha - \hat{\eta}\lfloor \alpha/\hat{\eta} \rfloor$ users from another profile. Next, for $r \in [P]$ and $m \in [\hat{\eta}]$, we consider the set $\mathcal{E}_r^m$ as:

$$\mathcal{E}_r^m = \bigcup_{i=0}^{\theta-1} \mathcal{Y}_r((i+m)\%\hat{\eta}), \tag{6}$$

where $\mathcal{Y}_r(i)$ is the i-th element of $\mathcal{Y}_r$. Indeed, $\mathcal{E}_r^m$ shifts $\mathcal{Y}_r$ to the right for m times, and picks θ elements from it. Then, for each $u \in \mathcal{E}_r^m$, we define the set $\mathcal{K}_r^{m,u}$ as follows.

$$\mathcal{K}_r^{m,u} = (u \| u)_{\nu_1} \| (f^* \| f^*)_{\nu_2 - \nu_1}, \tag{7}$$

where $\nu_1 = \binom{Q-2}{Q-\bar{t}-2}$ and $\nu_2 = \binom{Q-1}{Q-\bar{t}-1}$. Next, by defining $\mathcal{K}_r^{m,u}(s)$ as the s-th element of $\mathcal{K}_r^{m,u}$, for $s \in [\nu_2]$, it is assumed that $\mathcal{K}_{r,s}^{m,u}$ is the s circular shifts of $\mathcal{K}_r^{m,u}$, such that:

$$\mathcal{K}_{r,s}^{m,u} = \bigcup_{i=0}^{\nu_2} \mathcal{K}_r^{m,u}((i+s)\%\nu_2). \tag{8}$$

Moreover, we assume that $\overline{\mathcal{P}}_r = [P]_{\setminus r}$ for $r \in [P]$, and $\bar{\delta}_c = \overline{\mathcal{P}}_r(c)$ is the c-th element of $\overline{\mathcal{P}}_r$. The caching profiles in $\overline{\mathcal{P}}_r$ are sorted in descending order such that if $i < j$, then $\delta_{\overline{\mathcal{P}}_r(i)} \geq \delta_{\overline{\mathcal{P}}_r(j)}$. Next, for a given r and $c \in [P-Q+1]$, let $\mathcal{I}_c^r$ as:

$$\mathcal{I}_c^r = \big\{\mathcal{F} : \mathcal{F} \subseteq \big\{\overline{\mathcal{P}}_r(c+1), \cdots, \overline{\mathcal{P}}_r(P-1)\big\}, |\mathcal{F}| = Q-2\big\}. \tag{9}$$

Furthermore, denote the l-th $(Q-2)$-tuple of $\mathcal{I}_c^r$ by $\mathcal{I}_c^r(l)$, where $l \in \left[\binom{P-c-1}{Q-2}\right]$. For the delivery process with *Strategy B*, we use the so-called *transmission quintuple* (r, c, l, m, s), where $r \in [P]$, $c \in [P-Q+1]$, $l \in \left[\binom{P-c-1}{Q-2}\right]$, $m \in [\hat{\eta}]$ and $s \in [\nu_2]$. In each transmission quintuple (r, c, l, m, s), users assigned to the caching profiles $\mathcal{B} = \{\bar{\delta}_c\} \cup \mathcal{I}_c^r(l)$, and users in the set $\mathcal{E}_r^m$ are served. Suppose that $\mathcal{C} = \{\mathcal{F} : \mathcal{F} \subseteq \mathcal{B}, |\mathcal{F}| = \lfloor \alpha/\hat{\eta} \rfloor\}$, and $\mathcal{C}(n)$ is the n-th $\lfloor \alpha/\hat{\eta} \rfloor$-tuple of $\mathcal{C}$ with $n \in [\nu_2]$.

We define the function $I^+\big(\bar{\delta}_c, \mathcal{E}_r^m\big)$ such that $I^+\big(\bar{\delta}_c, \mathcal{E}_r^m\big) = 0$, if $\mathcal{E}_r^m = f^*$ and $\bar{\delta}_c = 0$; otherwise, $I^+\big(\bar{\delta}_c, \mathcal{E}_r^m\big) = 1$. If $I^+\big(\bar{\delta}_c, \mathcal{E}_r^m\big) = 1$, after eliminating the impacts of the phantom users f^*, the BS builds the transmission vector for the transmission quintuple (r, c, l, m, s) as follows.

$$\mathbf{x}_{r,c,l}^{m,s} = \sum\nolimits_{n=1}^{\nu_2} \sum_{k \in \mathcal{K}_{r,s}^{m,u}(n) \cup \mathcal{V}_p : u \in \mathcal{E}_r^m, p \in \mathcal{C}(n)} W_{\Theta_n, q}^k \mathbf{w}_{\mathcal{H}_{\mathcal{C}(n)}^k},$$

where $\Theta_n = \mathcal{B}_{\setminus \mathcal{C}(n)}$ with $|\Theta| = \bar{t}$, and $\mathbf{w}_{\mathcal{H}_{\mathcal{C}}(n)} \in \mathbb{C}^{L \times 1}$ is the precoder that suppresses the interference of user k at the set $\mathcal{H}_{\mathcal{C}(n)}^k = \{j \in \mathcal{E}_r^m \cup \mathcal{V}_p : p \in \mathcal{C}(n), j \neq k, f^*\}$. In [19, Appendix D], we prove that all users can decode their requested files with *Strategy B*. So, user i receives the signal

$$y_i = \mathbf{h}_i^{\mathrm{H}} \mathbf{x}_{r,c,l}^{m,s} + n_i.$$

In order to give further insight, an example for data delivery with *Strategy B* is provided in [19, Appendix C].

C. Unicast (UC) Data Delivery

In the UC delivery step, the BS transmits data to the users excluded from the CC delivery step. Here, unlike the CC delivery step that benefits from the global coded caching and spatial multiplexing gains, only local coded caching and spatial multiplexing gains are available. Suppose that the BS serves K_U users during the UC delivery step such that $K_U = \sum_{p=1}^{P} \max(0, \eta_p - \hat{\eta})$. Each of the requested files by these users is split into the same number of subpackets as in the CC delivery step. Then, in order to transmit these missing subpackets, we follow a greedy algorithm similar to [17], which comprises 3 processes: 1) sort users based on the number of their missing subpackets in descending order; 2) create a transmission vector to deliver one missing subpacket to each of the first $\min(\alpha, K_U)$ users; 3) repeat processes 1 and 2 until all missing files are transmitted.

IV. DoF Analysis

In this section, we use DoF as the metric of interest to measure performance. Here, the DoF is defined as the average number of users served concurrently during the delivery phase. In CC and UC delivery steps, we denote the total transmissions by T_M and T_U, respectively, and the number of served users by J_M and J_U. Therefore, DoF is computed as follows.

$$\mathrm{DoF} = \frac{J_M + J_U}{T_M + T_U}. \tag{10}$$

Furthermore, we suppose that K_M and K_U users are served during the CC and UC delivery steps such that $K_M = \sum_p \min(\hat{\eta}, \eta_p)$ and $K_U = \sum_p \max(0, \eta_p - \hat{\eta})$. The next theorem characterizes the DoF for the cache-aided networks operating with strategies *A* and *B* during the CC delivery step.

Theorem 1: Consider a dynamic MISO network with the spatial multiplexing gain of α, cache ratio γ and the delivery parameter $\hat{\eta}$. If the system operates with *Strategy A* in the CC delivery step, the DoF is given by:

$$\mathrm{DoF} = \begin{cases} \dfrac{K\binom{P-1}{Q-1}\beta}{\sum_{r=1}^{P-Q+1} D(\delta_r)\binom{P-r}{Q-1}} & K_U = 0 \\ \dfrac{K_M\binom{P-1}{Q-1}\beta + K_U(1-\gamma)\binom{P}{\bar{t}}\beta'}{\sum_{r=1}^{P-Q+1} D(\delta_r)\binom{P-r}{Q-1} + \left\lceil \frac{K_U(1-\gamma)\binom{P}{\bar{t}}\beta'}{\min(K_U,\alpha)} \right\rceil} & K_U \neq 0 \end{cases}, \tag{11}$$

where $\beta' = \beta\binom{P-\bar{t}-1}{Q-\bar{t}-1}$ and $D(\delta_r) = \phi_r$ if $\delta_r \neq 0$; otherwise, $D(\delta_r) = 0$. If *Strategy B* is applied during the CC delivery step, the DoF takes the form as follows.

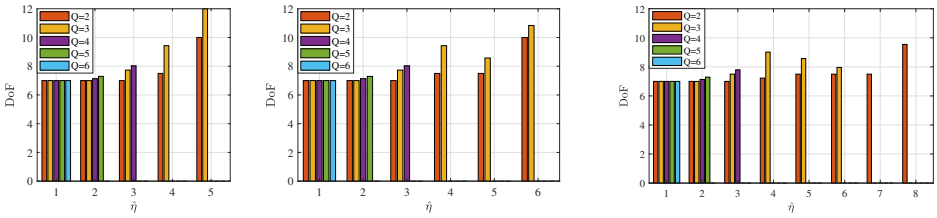
$$\mathrm{DoF} = \begin{cases} \dfrac{K\binom{P-1}{Q-1}(\hat{\eta}\bar{t}+\alpha)\nu_2}{N_M} & K_U = 0 \\ \dfrac{K_M\binom{P-1}{Q-1}(\hat{\eta}\bar{t}+\alpha)\nu_2 + K_U(1-\gamma)\binom{P}{\bar{t}}\alpha'}{N_M + N_U} & K_U \neq 0 \end{cases}, \tag{12}$$

where $\alpha' = (\hat{\eta}\bar{t} + \alpha)\nu_1\binom{P-\bar{t}-1}{Q-\bar{t}-1}$, $N_U = \left\lceil \frac{K_U(1-\gamma)\binom{P}{\bar{t}}\alpha'}{\min(K_U,\alpha)} \right\rceil$ and

$$N_M = \sum_{r=1}^{P} \sum_{c=1}^{P-Q+1} \sum_{m=1}^{\hat{\eta}} \sum_{s=1}^{\nu_2} \binom{P-c-1}{Q-2} I^+\big(\bar{\delta}_c, \mathcal{E}_r^m\big). \tag{13}$$

Proof: The proof is relegated to [19, Appendix D]. ∎

Remark 2: Suppose that K users are uniformly assigned to the caching profiles, i.e., $K = P\hat{\eta}$, and all users are served in the CC delivery step, i.e., $K_M = K$ and $K_U = 0$. If $\alpha \leq \hat{\eta}$, our schemes achieves the optimal DoF $\alpha(P\gamma + 1)$



(a) $\sigma = 0$ (b) $\sigma = 0.3$ (c) $\sigma = 3$

Fig. 2. The DoF versus $\hat{\eta}$ with different values of Q, $K = 30$, $\alpha = 7$, $P = 6$, $\gamma = \frac{1}{6}$ and $\bar{t} = 1$ for: (a) $\sigma = 0$, (b) $\sigma = 0.3$ and (c) $\sigma = 3$.

obtained in [16]. If $\alpha > \hat{\eta}$, the achievable DoF of our scheme is simplified to the optimal DoF $K\gamma + \alpha$ obtained in [15]. However, unlike [15], our scheme also supports the networks with non-integer $\frac{\alpha}{\hat{\eta}}$ (cf. [19, Appendix E]).

Indeed, increasing non-uniformness in user distribution prevents the system from achieving optimal DoF performance. For instance, for the case $\alpha > \hat{\eta}$, suppose that all users are served with *Strategy A* during the CC delivery step, such that $\beta = \hat{\eta}$ and $Q \leq \bar{t} + \lfloor \alpha/\hat{\eta} \rfloor$. In this setup, to boost the DoF performance, we can set $Q = \bar{t} + \lfloor \alpha/\hat{\eta} \rfloor$. By defining $\eta_{\text{avg}} = \frac{K}{P}$ and assuming α is divisible by $\hat{\eta}$ and η_{avg}, the DoF loss (compared to uniform user distribution) is $\alpha(1 - \eta_{\text{avg}}/\hat{\eta})$.

Setting $Q = \bar{t} + \lfloor \alpha/\hat{\eta} \rfloor$, however, requires to implement successive interference cancellation (SIC) at the receivers. To avoid using SIC, we can set $Q = \bar{t} + 1$, which simplifies the DoF of the uniform association to $\eta_{\text{avg}}(\bar{t} + 1)$. Here, we should compare the achievable DoF of $\eta_{\text{avg}}(\bar{t} + 1)$ with α such that: *i)* if $\eta_{\text{avg}}(\bar{t} + 1) \geq \alpha$, we use the proposed CC scheme to simultaneously benefit from global CC and spatial multiplexing gains; *ii)* if $\eta_{\text{avg}}(\bar{t} + 1) < \alpha$, we serve users via unicasting to not have any loss in DoF.

Remark 3: For the non-uniform user-to-profile association with $Q = \bar{t}+1$ and $\eta_1 \leq \alpha$, the best possible DoF is achievable by setting $\hat{\eta} = \eta_1$ (cf. [19, Appendix F]).

V. Numerical Results

In this section, we examine the impacts of non-uniformness in user distribution on the achievable DoF. In this regard, assume a cache-aided MISO setup with $\bar{t} = 1$, $P = 6$ and $\gamma = \frac{1}{6}$, in which $K = 30$ users are present during the delivery phase. Here, for each association, we compute the standard deviation σ as $\sigma^2 = \frac{1}{P}\sum_{p=1}^{P}(\eta_p - \eta_{\text{avg}})^2$, where $\eta_{\text{avg}} = 5$.

Fig. 2 illustrates the maximum achievable DoF for different Q and σ values with $\alpha = 7$. As observed from Fig. 2a, for the uniform user distribution, i.e., $\sigma = 0$, our scheme achieves the optimal DoF $K\gamma + \alpha = 12$ with $\hat{\eta} = \eta_{\text{avg}} = 5$ and $Q = \bar{t} + \lceil \alpha/\hat{\eta} \rceil = 3$. Here, we note that the system performance during the CC delivery step corresponds to *Strategy B* for the regime $\alpha > \hat{\eta}$ and non-integer $\alpha/\hat{\eta}$, which is missing in the literature. For small σ values (e.g., $\sigma = 0.3$), although the achievable DoF for $\hat{\eta} = 6$ and $Q = \bar{t} + 1 = 2$ is slightly less than $\hat{\eta} = 6$ and $Q = 3$, the receiver structure for $Q = 2$ is more straightforward, as they do not need to implement SIC. For large σ values (e.g., $\sigma = 3$), when $\max_p \eta_p > \alpha$, setting $Q = \bar{t} + 1$ and $\hat{\eta} = \max_p \eta_p$ maximizes the achievable DoF.

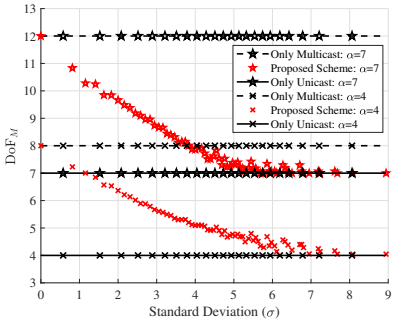


Fig. 3. The average of the maximum achievable DoF (DoF_M) versus the standard deviation (σ) with $K = 30$, $\gamma = \frac{1}{6}$, $P = 6$, $\bar{t} = 1$ and $\alpha \in \{4, 7\}$.

In Fig. 3, for any association, we find $\text{DoF}_{\max}$ which is the maximum achievable DoF obtained by a line search over $\hat{\eta}$ and Q values. Accordingly, for the associations with the same σ value, DoF_M indicates the average of $\text{DoF}_{\max}$. Here, our proposed scheme is compared with the optimal case, which corresponds to uniform user distribution (described as *only multicast*), and the case that all users are served via unicasting (described as *only unicast*). Although the placement phase was designed solely based on γ, our scheme boosts the maximum achievable DoF by $10\% - 70\%$ over unicasting for moderate σ (e.g., $\sigma = 1 - 4.5$). So, to maximize the achievable DoF, the server should serve users via the proposed approach for moderate σ, and serve users via unicasting for large σ.

VI. Conclusion

We proposed a novel coded caching scheme for handling network dynamicity where the users can freely enter or depart the network at any time. The conventional schemes in the literature are not truly dynamic as they are only applicable if: *1)* minimum profile length (the number of users assigned to the profile) exceeds the spatial multiplexing gain α, and *2)* $\alpha \geq \hat{\eta}$ and α exceeds the global CC gain, where $\hat{\eta}$ can be the length of any profile. Our proposed scheme addressed this bottleneck by providing a universal solution applicable to any dynamic network setup, removing all the constraints imposed by existing solutions. We also analyzed the degrees-of-freedom (DoF) performance of the proposed scheme, and for the uniform distribution, we showed that it achieves the optimal DoF not only in the regions covered in the literature but also in the region $\alpha \geq \hat{\eta}$ with non-integer $\alpha/\hat{\eta}$.

REFERENCES

[1] E. Summary, "Cisco Visual Networking Index – Forecast and," *Europe*, vol. 1, pp. 2007–2012, 2012.

[2] H. B. Mahmoodi, M. J. Salehi, and A. Tolli, "Non-Symmetric Coded Caching for Location-Dependent Content Delivery," in *IEEE International Symposium on Information Theory (ISIT)*, July 2021, pp. 712–717.

[3] M. Salehi, K. Hooli, J. Hulkkonen, and A. Tolli, "Enhancing Next-Generation Extended Reality Applications with Coded Caching," *arXiv preprint arXiv:2202.06814*, 2022. [Online]. Available: http://arxiv.org/abs/2202.06814

[4] M. A. Maddah-Ali and U. Niesen, "Fundamental Limits of Caching," *IEEE Transactions on Information Theory*, vol. 60, no. 5, pp. 2856–2867, 2014.

[5] S. P. Shariatpanahi, G. Caire, and B. H. Khalaj, "Multi-antenna Coded Caching," in *IEEE International Symposium on Information Theory (ISIT)*, June 2017, pp. 2113–2117.

[6] J. Zhao, M. M. Amiri, and D. Gunduz, "Multi-antenna Coded Content Delivery with Caching: A Low-Complexity Solution," *IEEE Transactions on Wireless Communications*, vol. 19, no. 11, pp. 7484–7497, 2020.

[7] S. P. Shariatpanahi, S. A. Motahari, and B. H. Khalaj, "Multi-server Coded Caching," *IEEE Transactions on Information Theory*, vol. 62, no. 12, pp. 7253–7271, 2016.

[8] A. Tölli, S. P. Shariatpanahi, J. Kaleva, and B. H. Khalaj, "Multi-antenna Interference Management for Coded Caching," *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 2091–2106, 2020.

[9] S. P. Shariatpanahi, G. Caire, and B. Hossein Khalaj, "Physical-Layer Schemes for Wireless Coded Caching," *IEEE Transactions on Information Theory*, vol. 65, no. 5, pp. 2792–2807, 2019.

[10] E. Lampiris and P. Elia, "Bridging two extremes: Multi-antenna Coded Caching with Reduced Subpacketization and CSIT," in *IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, July 2019, pp. 1–5.

[11] M. Salehi, A. Tolli, S. P. Shariatpanahi, and J. Kaleva, "Subpacketization-Rate Trade-Off in Multi-Antenna Coded Caching," in *IEEE Global Communications Conference (GLOBECOM)*, December 2019, pp. 1–6.

[12] M. J. Salehi, E. Parrinello, S. P. Shariatpanahi, P. Elia, and A. Tolli, "Low-Complexity High-Performance Cyclic Caching for Large MISO Systems," *IEEE Transactions on Wireless Communications*, vol. 21, no. 5, pp. 3263 – 3278, 2021.

[13] Q. Yan, M. Cheng, X. Tang, and Q. Chen, "On the Placement Delivery Array Design for Centralized Coded Caching Scheme," *IEEE Transactions on Information Theory*, vol. 63, no. 9, pp. 5821–5833, 2017.

[14] 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 Transactions on Information Theory*, vol. 67, no. 9, pp. 6193 – 6222, 2021.

[15] E. Parrinello, P. Elia, and E. Lampiris, "Extending the Optimality Range of Multi-Antenna Coded Caching with Shared Caches," in *IEEE International Symposium on Information Theory (ISIT)*, June 2020, pp. 1675–1680.

[16] E. Parrinello, A. Ünsal, and P. Elia, "Fundamental Limits of Coded Caching with Multiple Antennas, Shared Caches and Uncoded Prefetching," *IEEE Transactions on Information Theory*, vol. 66, no. 4, pp. 2252 – 2268, 2019.

[17] M. Abolpour, M. J. Salehi, and A. Tolli, "Coded Caching and Spatial Multiplexing Gain Trade-off in Dynamic MISO Networks," in *IEEE Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, July 2022, pp. 1–5.

[18] M. Salehi, E. Parrinello, H. B. Mahmoodi, and A. Tolli, "Low-Subpacketization Multi-Antenna Coded Caching for Dynamic Networks," in *Joint European Conference on Networks and Communications & 6G Summit (EuCNC/6G Summit)*, June 2022, pp. 112–117.

[19] M. Abolpour, M. Salehi, and A. Tölli, "Cache-Aided Communications in MISO Networks with Dynamic User Behavior: A Universal Solution," *arXiv preprint arXiv:2304.11623*, 2023. [Online]. Available: https://arxiv.org/abs/2304.11623