Amidzade, Mohsen; Tirkkonen, Olav; Caire, Giuseppe

Optimal Multicast-Cache-Aided On-demand Streaming in Heterogeneous Wireless Networks via a Path/Surface Following Approach

Published in:
IEEE Transactions on Wireless Communications

DOI:
10.1109/TWC.2023.3345223

Published: 01/07/2024

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Amidzade, M., Tirkkonen, O., & Caire, G. (2024). Optimal Multicast-Cache-Aided On-demand Streaming in Heterogeneous Wireless Networks via a Path/Surface Following Approach. *IEEE Transactions on Wireless Communications*, *23*(7), 7833-7848. https://doi.org/10.1109/TWC.2023.3345223

Optimal Multicast-Cache-Aided On-demand Streaming in Heterogeneous Wireless Networks via a Path/Surface Following Approach

Mohsen Amidzadeh*, Olav Tirkkonen* and Giuseppe Caire†
*Department of Communications and Networking, Aalto University, Finland
†Communications and Information Theory Chair, TU Berlin, Germany

***Abstract*—We consider a hybrid streaming scheme based on cache-enabled orthogonal multipoint multicast (OMPMC) and on-demand single-point unicast (SPUC) transmission. The network contains two types of nodes, cache-equipped helper nodes (HNs) handling content-centric OMPMC, and cellular base stations (BSs) handling user-centric SPUC. The OMPMC service streams cached files across the network. Users whose demands cannot be satisfied by OMPMC, either because of poor signal quality or because the requested file is not cached at HNs, are served by SPUC; requested files are fetched from the core network and unicast to users using group-specific beamforming transmissions. We consider the overall network radio resource consumption to satisfy the users' requests for a given outage probability. This yields a parametric constrained optimization problem over the cache and resource allocations of the OMPMC component, as well as the multi-user beamforming scheme of the SPUC component. We devise a surface-following approach on the basis of path-following method to find the optimal traffic streaming solution. Simulation results show that the hybrid scheme provides a more promising trade-off between resource consumption and service outage probability, compared to OMPMC-only and SPUC-only alternatives.**

***Index Terms*—Hybrid content streaming, wireless caching, multipoint multicasting, single-point unicasting, user group-specific beamforming, surface following approach.**

I. Introduction

Wireless caching [1] is a promising approach to mitigate the unprecedented data congestion and traffic escalation issues in cellular networks. To achieve a viable cache strategy, the two phases of cache placement and cache delivery need to be considered [2], and both should be optimized.

Content placement can be performed using a probabilistic [3]–[6], or a deterministic approach [2], [7], [8], where the probabilistic one can be applied to large networks in a scalable manner. In [3], a probabilistic content placement was proposed by which the cache-equipped nodes independently and randomly store files. In [4], [5], a tier-specific probabilistic caching was exploited for the Heterogeneous Networks (HetNets). In [6], a two-tier HetNet was considered with a hybrid cache placement, combining deterministic caching in one tier, and probabilistic caching in the other, was devised.

This work was funded in part by the Academy of Finland (grant 345109). The work of Giuseppe Caire was partially funded by the European Research Council under the ERC Advanced Grant N. 789190, CARENET.

When it comes to the content delivery, multipoint multicast (MPMC), single-point multicast (SPMC), and single-point unicast (SPUC) schemes should be distinguished. MPMC utilizes multiple serving nodes to cooperatively broadcast files across the network, and as such it establishes a content-centric scheme. However, SPUC leverages on-demand transmissions to individually serve User Equipments (UEs), and is considered as a UE-centric scheme. For SPMC, each base station (BS) individually multicasts a file to several requesting UEs.

SPUC is the main content delivery method underlying the success of mobile and local area networks. Mobile networks, operating in licensed frequency bands, provide the highest capacity when operating with frequency reuse 1. This was first understood in the context of 3rd generation CDMA networks, see [9], and has been verified for 4G networks [10] and is the underlying idea of massive-MIMO, where pilot sequences may have a higher reuse factor, but data transmissions operate with reuse 1 [11]. SPUC has been considered in [4], [7], [12]–[15] as an on-demand cache delivery scheme for HetNets. In [4], zero forcing (ZF) beamforming BSs and cache-equipped helper nodes were considered. In [12], a tier-level resource allocation was utilized with SPUC to optimize a cache policy from the successful offloading probability perspective. An SPUC with a deterministic UE association mechanism has been applied in [14] to optimize the total download latency in a HetNet applying UE-specific orthogonal resources.

SPMC has been used in [16]–[18] as a content delivery method. In [17], coded caching is applied for a cellular network with multi-antenna BSs. The authors exploit joint SPUC and SPMC beamforming for content delivery, and beamforming vectors are optimized to maximize the minimum rate of the UEs. In [18], the authors apply a delivery scheme combining SPUC and SPMC for a cache-enabled HetNet. The multicast part is scheduled to optimize the energy consumption of the network. In [19], beamforming multicast is used to develop a cache policy that optimizes the overall transmission delay, taking into account the beamforming parameters.

Digital Terrestrial TV Broadcasting systems [20] deliver content over multipoint broadcast transmissions, while multipoint multicast (MPMC) delivery is exploited in the Long Term Evolution (LTE) system in the context of the enhanced multimedia broadcast–multicast service (eMBMS) [21], [22]. In a multi-cell transmission mode, all serving BSs stream the

same file through the whole network in a Single-Frequency-Network (SFN) configuration. Therefore, each file is concurrently transmitted over the same frequency bandwidth. SFN-configured orthogonal MPMC (OMPMC) was utilized in [23] for edge caching cellular networks, where network-wide resources are orthogonalized among cached files.

In this paper[1], we devise a hybrid streaming scheme for cache-enabled multi-antenna HetNets, combining OMPMC and SPUC. SPUC individually serves UEs, but suffers from co-channel interference (CCI). In contrast, OMPMC serves a population of UEs interested in a given file with limited CCI [25]. This portrays a trade-off between OMPMC and SPUC schemes. We thus exploit this contrast to develop a hybrid scheme benefiting from the advantages of both.

Contrary to [17], [26], we devise the hybrid scheme such that all UEs being dissatisfied by the multicast component are properly served by the network. In [17], [18], shared resources are used for the joint SPUC and SPMC transmissions. In contrast, here, we use disjoint resources for OMPMC and SPUC. In [26] the outage probability of the network was optimized regardless of the resource consumed by the whole network. However, we here optimize the network with respect to the total amount of consumed resources subject to inevitable outages due to coping with the possibility of too large bandwidth requests. In [16], [18], [27], SPMC networking is used for cache delivery. In contrast to SPMC which suffers from the CCI, we here utilize OMPMC, as a content-centric networking component, which can considerably reduce CCI, exploit a multipoint transmission mechanism and consequently increase the network performance. Furthermore, we have incorporated a beamforming SPUC scheme as a UE-centric networking component to individually serve UEs whose requests cannot be satisfied by content-centric networking.

The main contributions of this paper are as follows.

- We develop a hybrid scheme, based on content-centric OMPMC and UE-centric beamforming SPUC, for traffic streaming policy in edge-caching multi-antenna HetNets.
- Using stochastic geometry, we formulate the problem of optimal traffic offloading as a joint optimization problem over cache placement, resource allocation and multiuser beamforming. We then draw some properties of this problem and characterize its global optimum. To solve this problem, we represent a multi-parametric optimization problem and devise a surface-following approach on the basis of path-following method to find the globally optimal policy.
- We compare the proposed scheme to the baseline approaches: OMPMC-only and SPUC-only, and numerically demonstrate the advantage of the hybrid approach. The intuition behind our results is that popular files should be cached in the network in order to offload as much traffic as possible with OMPMC, while the less popular files, together with requests of users in very bad OMPMC reception conditions, are served by cellular unicast.

A possible practical consequence of this study is the following: In today's systems, multicast in the form of eMBMS is reserved for "live streaming" (e.g., live TV channels), while on-demand streaming is uniquely handled by unicast. In contrast, our results suggest that it is convenient to use eMBMS (combined with harmonic broadcasting [28]) for *on-demand streaming* of popular content, while handling specific (possibly unpopular) user requests by standard unicast. The approach of this paper can be applied on top of contemporary and future mobile networks which have both an eMBMS and a conventional SPUC component. We show that by employing this approach, the network can take advantage of a delivery scheme that offers an improved contrast between quality-of-service and total resource consumption.

[1]Early results of this work were presented in [24].

The remainder of this paper is organized as follows. In Section II, the system model is given. In Section III, the resource consumption of hybrid scheme is computed. The problem is analyzed in Section V, and simulation results are discussed in Section VI. Section VII concludes the paper.

Notations: We use lower-case a for scalars, bold-face lower-case $\mathbf{a}$ for vectors and bold-face uppercase $\mathbf{A}$ for matrices. $\mathbf{A}^\top$ and $\mathbf{A}^\dagger$ are the transpose and Hermitian conjugate of $\mathbf{A}$ and $\|\mathbf{a}\|$ is the Euclidean norm of $\mathbf{a}$. The notation $[a_{ij}]_{i,j}$ shows a matrix whose i-th row and j-th column is a_{ij} and $[a_i]_i$ is a column vector with i-th component being a_i. The notation $\mathbf{a}^{-1}$ denotes the component-wise inversion of $\mathbf{a}$, and $\mathbf{1}$ and $\mathbf{0}$ to denote the vector with all elements equal to one and zero, respectively. We use $\dot{\boldsymbol{a}}(\theta)$ to represent the derivative of $\boldsymbol{a}(\theta)$ with respect to θ. Further, δ_{ij} is the Kronecker delta function, $\mathbb{C}$ and $\mathbb{R}$ denote the complex valued and real valued numbers.

II. System Model

We consider a content library with N different files. There is a set of active UEs, and in a given time-slot, each UE is interested in one of the files. The fraction of UEs preferring file n is f_n and is called the file popularity. For the simulation, the Zipf distribution [29] is used as the file popularity; the applied methodology and analysis are not restricted to this. We consider the same streaming rate R for all files. The study of dynamic rate is left for future work.

We consider a two-tier HetNet with BSs and helper nodes (HNs). The network applies a hybrid delivery scheme, based on OMPMC and SPUC components to serve UEs as shown in Fig. 1. The HNs constitute the multicast component and are equipped with limited-capacity caches which store files based on a probabilistic approach [30]. The HNs apply OMPMC to **proactively** broadcast the cached files through the whole network in file-specific disjoint resources. Note that in contrast to on-demand multicast schemes [16], [27], the OMPMC approach multicasts the files in a proactive manner with respect to the requests of UEs. As such, the UEs are not served by a specific HN, but are associated with the entire multicast component. In the OMPMC layer, there may be CCI arising from far-away HNs. However, with conventional orthogonal frequency division multiplexing (OFDM) settings, such CCI would be negligible [25]. The BSs, on the other hand, constitute the unicast component and are connected to a high-capacity backhaul link which can **on-demand** fetch files from the core-network, and transmit these to requesting UEs using an unicast transmission. In contrast to the OMPMC multicast

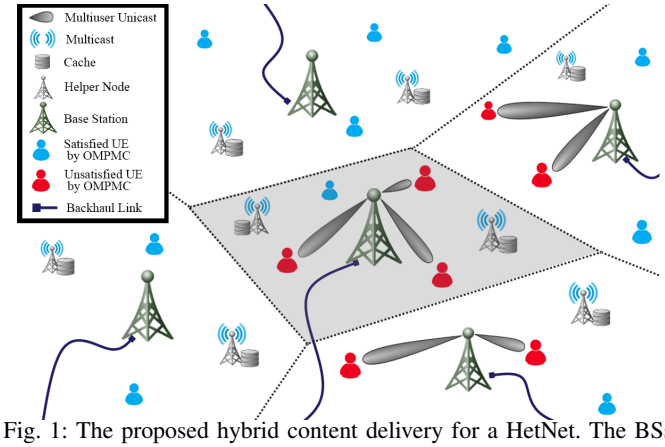

Fig. 1: The proposed hybrid content delivery for a HetNet. The BSs constitute a Voronoi tessellation that is specified with dashed lines.

component, it individually responds to those UEs that are not satisfied by OMPMC proactive service. The BSs are equipped with L antennas and exploit multiuser ZF beamforming to deliver the contents towards requesting UEs. Notice that each requesting UE of this layer is associated with its nearest BS, and as such BS-specific association applies in contrast to the tier-specific association of the multicast layer.

The network dedicates two network-wide disjoint resources for OMPMC and SPUC, denoted by W^{MC} and W^{UC}, respectively. Each BS uses the full bandwidth W^{UC}, i.e., frequency reuse 1 is used in the unicast layer [9]–[11]. No time synchronization is needed between OMPMC and SPUC to properly cooperate in hybrid delivery. Specifically, OMPMC continually broadcasts the cached files in a time-slotted fashion, and each UE dissatisfied by OMPMC is associated with the nearest BS to be on-demand served by the SPUC component. The only requirement is that the radio resources of these schemes remain disjoint during the network operation.

The locations of UEs, BSs and HNs are assumed based on three independent Poisson Point Processes (PPPs), Φ_{ue}, Φ_{bs} and Φ_{hn}, respectively and with intensities λ_{ue}, λ_{bs} and λ_{hn}.

A. Cache Placement and Content Fetching

The HNs are equipped with caches of limited capacity so that they can store at most M files. They independently cache files based on a probabilistic placement policy [3]. Specifically, we assume that the network operates on file chunks of equal size. To cache file n, each HN refers to a network-wide caching weight $\{p_n\}_{n=1}^{N}$, $0 \leq p_n \leq 1$, where p_n indicates the probability that file n is stored at a randomly selected HN. To comply with the cache capacity of HNs, we have $\sum_{n=1}^{N} p_n \leq M$. To store the files, each HN first creates an arrangement with M unit-length segments. As depicted in Figure 2, it then fills the segments based on the values of $\{p_n\}_{n=1}^{N}$, and proceeds with filling in the next segment if not enough space is available in the current segment for a weight. Each of the M segments thus have a set of possible files that can be cached in the corresponding part of the cache memory. At the end, when the segments are completely filled, the HN generates a uniform random variable $\mathcal{U} \in [0, 1]$, and stores files being intersected with this random variable. Note that the

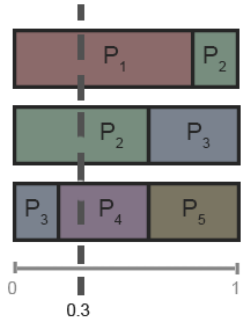

Fig. 2: An example of probabilistic cache placement with $N = 5$, $L = 3$, $\boldsymbol{p} = [\frac{4}{5}, \frac{4}{5}, \frac{3}{5}, \frac{2}{5}, \frac{2}{5}]$, $\mathcal{U} = 0.3$. Only first, second and fourth files will be cached.

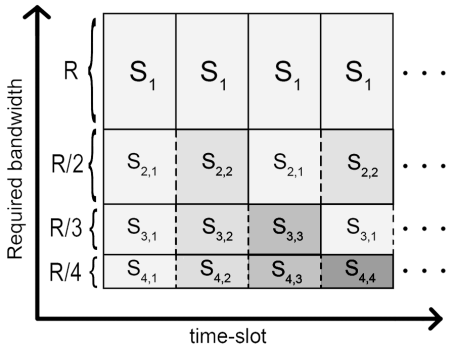

Fig. 3: An example of harmonic broadcasting protocol with $D_{\mathrm{hb}} = 4$.

files cached at HNs follow a PPP with intensity $\{\lambda_{\mathrm{ue}} p_n\}_{n=1}^{N}$, according to the thinning property of Φ_{ue} [31].

The BSs providing the on-demand unicast delivery, reactively fetch the requested files from the core network, if they are requested by UEs.

B. Content Delivery and Group-Specific Beamforming

Within each time-slot of network operation, there exists a spatial realization of interested UEs based on PPP Φ_{ue}. The network applies a hybrid delivery scheme, based on **content-centric** OMPMC and **UE-centric** SPUC, to serve UEs. Any UE not being satisfied by the multicast component is assigned to the unicast component to be properly served.

The multicast component proactively streams cached files across the network without any requests. For this, it applies the OMPMC scheme on HNs. Therefore, a UE preferring file n can receive it based on an aggregation of received signals from all HNs caching file n. Moreover, the OMPMC transmissions continuously repeat on a time-slot basis. To remove the co-channel interference, it uses file-specific disjoint resources for broadcasting different files with w_n^{MC} being the bandwidth allocated for file n.

To reduce the streaming latency in the OMPMC mode, we apply the Harmonic Broadcasting (HB) protocol [28], which reduces the latency by increasing the transmission bandwidth by an efficient approach. This protocol enables UEs to download a file from the beginning with a considerably lower latency than to the case where the file would be streamed in its entirety in a time division multiplexed manner. Then, a requesting UE would have to wait for the beginning of next file delivery. The HB protocol performs as follows:

- Each cached file, with streaming rate R, is equally divided into $D_{\rm hb}$ segments: $\{s_i\}_{i=1}^{D_{\rm hb}}$.
- Each i-th segment is again divided into i sub-segments: $\{s_{i,j}\}_{j=1}^{i}$ and is transmitted by the information rate R/i.
- Different segments $\{s_i\}_{i=1}^{D_{\rm hb}}$ are continuously broadcasted in the disjoint resources they require.

Figure 3 shows an example of HB protocol with $D_{\rm hb} = 4$. Considering that the bandwidth of i-th segment is R/i, the total bandwidth consumption for a file, with original streaming rate R, is increased by a factor of $N_{\rm hb} = \sum_{i=1}^{D_{\rm hb}} 1/i$, where $N_{\rm hb}$ is a harmonic number. The OMPMC repeatedly broadcasts segments of each cached file in segment-specific resources by applying the HB protocol. The time-slot operation of the network now happens with duration reduced by a factor of $1/D_{\rm hb}$. To receive a file from OMPMC, it needs to wait for the beginning of the next $1/D_{\rm hb}$ fraction of the time-slot, instead of waiting for the next time-slot of file transmission. For a file with a length of S seconds, this protocol notably reduces the average streaming latency from $S/2$ to $S/2D_{\rm hb}$ seconds. As an instance, the average latency can be decreased by a factor of $D_{\rm hb} = 226$, if the harmonic number is set to $N_{\rm hb} = 6$, as $\sum_{i=1}^{226} 1/i = 6$. As a consequence, the required bandwidth to harmonic-broadcast a file with desired rate R is increased to $6R$.

We assume that the average transmission power of all HNs is the same, denoted by $p_{\rm tx}$. Further, each HN allocates a fractional power $p_{\rm tx} w_n^{\rm MC}/W^{\rm MC}$ to broadcast file n in resource $w_n^{\rm MC}$, where $W^{\rm MC} = \sum_{n=1}^{N} w_n^{\rm MC}$ is the total resource for the multicast component. In this case, the transmission Signal-to-Noise-Ratio (SNR) related to file n is

$$\gamma_{n,\rm tx}^{\rm MC} = \frac{p_{\rm tx} w_n^{\rm MC}/W^{\rm MC}}{w_n^{\rm MC} N_0} = \frac{p_{\rm tx}}{W^{\rm MC} N_0} \;:=\; \gamma_{\rm tx}^{\rm MC}, \qquad (1)$$

where N_0 is the noise spectral density. Note that the Tx-SNR is independent of file index n.

Any requesting UE not being satisfied by the multicast component, due to the file not being cached or poor SNR in the OMPMC component, requests the file from the unicast component. As such, the BSs in the unicast layer, constitute a Poisson-Voronoi tessellation with different cell sizes. The nearest BS to the UE fetches the file from the core network and unicasts it towards the UE. If the Signal-to-Interference-plus-Noise Ratio (SINR) of a user is above a threshold, the responsible BS responds to the UE by allocating the resources it needs to successfully decode its files. In each Voronoi cell, there may be U requesting UEs associated with the responsible BS, where U is a random variable depending on $\Phi_{\rm ue}$ and $\Phi_{\rm bs}$. Let $w_k^{\rm UC}$ denote the amount of radio resources needed for UE k to be successfully served. The BS applies a ZF beamforming to simultaneously serve n_z UEs at the same frequency resource, where n_z is a beamforming parameter controlled by the unicast component. To serve all U UEs requiring resources $\{w_k^{\rm UC}\}_{k=1}^{U}$ in the cell, the BS performs the following group-specific policy. It first clusters the UEs into U/n_z groups. Then, it simultaneously transmits to n_z UEs of a specific group by using ZF beamforming in a radio resource equaling to the maximum value of resources needed by UEs in the group. It can then group-wise serve all UEs using group-specific ZF beamformings characterized by parameter n_z. It is noteworthy that the average transmission power of all BSs is set to $p_{\rm tx}$.

III. Resource Consumption Analysis

A. Resource Consumption of Multicast Component

The multicast component applies network-wide resource allocation with OFDM-based transmission to broadcast the files through the network. With synchronous transmissions of the BSs, the signal-to-noise-ratio (SNR) of UE k requesting file n is expressed as [23]:

$$\gamma_{k,n}^{\rm MC} = \gamma_{\rm tx}^{\rm MC} \sum_{j \in \Phi_{\rm hn,n}} |h_{j,k}|^2 \|\boldsymbol{x}_k - \boldsymbol{r}_j\|^{-e}, \qquad (2)$$

where $\gamma_{\rm tx}^{\rm MC}$ is the Tx-SNR from (1), $\Phi_{\rm hn,n}$ stands for the set of HNs caching file n, $h_{j,k}$ is the channel coefficient between HN j and UE k, $\boldsymbol{x}_k$ and $\boldsymbol{r}_j$ are the locations of UE k and HN j, respectively, and e is the path-loss exponent.

We use a standard distance-dependent path-loss model, and assume a Rayleigh distribution for the channel coefficient, i.e., $|h_{j,k}|^2 \sim \exp(1)$. We assume quasi-static fading within a slot. Given the SNR $\gamma_{k,n}^{\rm MC}$, the maximum achievable rate for a transmission is obtained from the Additive-White-Gaussian-Noise channel capacity. For a broadcast transmission with streaming rate R, this translates to an outage probability for user k receiving file n; If the maximum rate is less than the R, the UE is in outage. The outage probability $\mathcal{O}_{n,k}^{\rm MC}$ for UE k requesting file n thus is:

$$\mathcal{O}_{n,k}^{\rm MC} = \mathbb{P}\{w_n^{\rm MC} \log_2(1 + \gamma_{k,n}^{\rm MC}) \le R\}.$$

We define a spectral efficiency threshold α_n and a resource weight β_n as:

$$\alpha_n = \frac{1}{\beta_n} := \frac{R}{w_n^{\rm MC}}, \qquad (3)$$

where β_n is the radio resource normalized by streaming rate R, which is allocated by OMPMC to broadcast file n. Note that α_n characterizes the modulation and coding scheme jointly applied by all the caching HNs when transmitting file n. The file is thus in outage at user k if:

$$\gamma_{k,n}^{\rm MC} \le 2^{\alpha_n} - 1\,. \qquad (4)$$

The total resource usage of the multicast component then is $W^{\rm MC} = \sum_{n=1}^{N} w_n^{\rm MC} = R \sum_{n=1}^{N} \beta_n$.

As we apply HB protocol, the effective resource consumption will be multiplied by $N_{\rm hb}$, i.e.,

$$W_{\rm eff}^{\rm MC}(\boldsymbol{\alpha}) = N_{\rm hb}\, R \sum_{n=1}^{N} \beta_n, \qquad (5)$$

where $\boldsymbol{\alpha} = [\alpha_i]_i$. Based on the Slivnyak-Mecke theorem [31], the performance can be computed for a typical UE located at the origin. Hence, we can set $\mathcal{O}_{n,0}^{\rm MC} = \mathcal{O}_{n}^{\rm MC}$.

Accordingly, the outage probability of file n, served by the multicast component, is [23]:

$$\mathcal{O}_n^{\mathrm{MC}}(p_n,\boldsymbol{\alpha}) = \frac{2}{\pi}\int_0^\infty \left\{ \frac{1}{w}\cos\left(\frac{\pi^2\lambda_{\mathrm{hn}}p_n}{e\cos(\frac{\pi}{e})}\left(\frac{w}{\widetilde{\alpha}_n}\right)^{2/e}\right) \times \exp\left(\frac{-\pi^2\lambda_{\mathrm{hn}}p_n}{e\sin(\frac{\pi}{e})}\left(\frac{w}{\widetilde{\alpha}_n}\right)^{2/e}\right)\sin\left(\frac{w}{\gamma_R}\right)\right\} dw, \quad (6)$$

where $\widetilde{\alpha}_n = \left(2^{\alpha_n}-1\right)\sum_{l=1}^{N}\beta_l$ and $\gamma_R = \dfrac{p_{\mathrm{tx}}}{N_0}\dfrac{1}{R\,N_{\mathrm{hb}}}$. It can be verified that as p_n or β_n grow, the outage probability decreases for different path-loss exponent $e>2$. Based on (6), the HB protocol worsens the multicast outage such that the resulting deficiency can be compensated if HN intensity λ_{hn} is multiplied by a factor of $N_{\mathrm{hb}}^{2/e}$. For the path-loss exponent $e=4$, as a typical value for moderate and large distances [32], a closed-form expression is obtained:

$$\mathcal{O}_n^{\mathrm{MC}}(p_n,\boldsymbol{\alpha}) = \mathrm{erfc}\left(\frac{\pi^2\lambda_{\mathrm{hn}}p_n}{4}\sqrt{\frac{\gamma_R}{\widetilde{\alpha}_n}}\right).$$

B. Resource Consumption of Unicast Component

Following the trend of contemporary mobile networks [9]–[11], we assume reuse 1 in the SPUC component. In each Voronoi cell of the unicast layer, there may be U UEs requesting from the BS of the cell. These UEs are grouped, and group-specific ZF beamformings are utilized such that each ZF beamforming simultaneously serve n_z UEs in the same frequency bandwidth. The transmitted signal of BS j beamforming towards n_z UE is $\boldsymbol{q}_j = \sqrt{p'_{\mathrm{tx}}}\sum_{i=1}^{n_z}\boldsymbol{v}_j^i s_j^i$, where $\boldsymbol{v}_j^i \in \mathbb{C}^{L\times 1}$ and $s_j^i \in \mathbb{C}$ are the unit-norm ZF precoding vector and the data symbol of BS j aiming for UE i, respectively. Having unit-norm precoding vectors, we need to have $p'_{\mathrm{tx}} = p_{\mathrm{tx}}/n_z$ to comply with the transmission power constraint of BSs. As such, the received signal of UE k being served by it nearest BS (here called BS 0 with location $\boldsymbol{r}_0$) is expressed as:

$$\begin{aligned} y_k =& \sqrt{p'_{\mathrm{tx}}}\sum_{j\in\Phi_{\mathrm{bs}}}\sum_{i=1}^{n_z}\boldsymbol{h}_j^{k\dagger}\boldsymbol{v}_j^i\|\boldsymbol{x}_k-\boldsymbol{r}_j\|^{-e/2}s_j^i \;+\; n_k \\ =& \underbrace{\sqrt{p'_{\mathrm{tx}}}\,\boldsymbol{h}_0^{k\dagger}\boldsymbol{v}_0^k\|\boldsymbol{x}_k-\boldsymbol{r}_0\|^{-e/2}s_0^k}_{\text{desired signal}} + \underbrace{\sqrt{p'_{\mathrm{tx}}}\sum_{i\neq k}^{n_z}\boldsymbol{h}_0^{k\dagger}\boldsymbol{v}_0^i\|\boldsymbol{x}_k-\boldsymbol{r}_0\|^{-e/2}s_0^i}_{\text{intra-cell interference}} \\ &+ \underbrace{\sqrt{p'_{\mathrm{tx}}}\sum_{j\in\Phi_{\mathrm{bs}}\setminus\{0\}}\sum_{i=1}^{n_z}\boldsymbol{h}_j^{k\dagger}\boldsymbol{v}_j^i\|\boldsymbol{x}_k-\boldsymbol{r}_j\|^{-e/2}s_j^i}_{\text{inter-cell interference}} \;+\; n_k, \end{aligned}$$

where $\boldsymbol{h}_j^k \in \mathbb{C}^{L\times 1}$ is the channel coefficient between BS j and UE k with complex Gaussian distribution, n_k is the complex Gaussian noise at the UE with variance N_0W^{UC}, and $\boldsymbol{r}_0$ is the location of the nearest BS to the UE.

For ZF beamforming, the precoding vectors $\{\boldsymbol{v}_0^i\}_{i=1}^{n_z}$ of the BS of interest are chosen such that the intra-cell interference is perfectly removed. As such, $\boldsymbol{v}_0^k$ is set as the k-th column of matrix $\boldsymbol{H}_0\left(\boldsymbol{H}_0^\dagger\boldsymbol{H}_0\right)^{-1}$, where $\boldsymbol{H}_0 = \left(\boldsymbol{h}_0^1,\ldots,\boldsymbol{h}_0^{n_z}\right)$. Consequently $\boldsymbol{h}_0^{k\dagger}\boldsymbol{v}_0^i = 0$ for $i\neq k$ and for UE k served by the unicast component, the SINR is:

$$\gamma_k^{\mathrm{UC}} = \frac{g_0^k\|\boldsymbol{x}_k-\boldsymbol{r}_0\|^{-e}}{1/\gamma_{\mathrm{tx}}^{\mathrm{UC}} + \underbrace{\sum_{j\in\Phi_{\mathrm{bs}}\setminus\{0\}} g_j^k\|\boldsymbol{x}_k-\boldsymbol{r}_j\|^{-e}}_{I_k}}, \quad (7)$$

where the unicast Tx-SNR is $\gamma_{\mathrm{tx}}^{\mathrm{UC}} = p_{\mathrm{tx}}/(n_zW^{\mathrm{UC}}N_0)$. Note that $g_0^k = \left|\boldsymbol{h}_0^{k\dagger}\boldsymbol{v}_0^k\right|^2 \sim \Gamma(L-n_z+1,1)$ [33], where g_0^k is the effective channel coefficient between nearest BS and UE k and $\Gamma(a,b)$ is the gamma distribution with shape a and scale b. Further, $g_j^k = \sum_{i=1}^{n_z}\left|\boldsymbol{h}_j^{k\dagger}\boldsymbol{v}_j^i\right|^2$ is the effective channel coefficient from BS j to UE k. Considering that for $j\neq 0$, vectors $\boldsymbol{h}_j^k$ and $\boldsymbol{v}_j^i$ are independent and the correlation among variables $\left\{\boldsymbol{h}_j^{k\dagger}\boldsymbol{v}_j^i\right\}_{i=1}^{n_z}$ for different values of i, can be neglected, we thus get $g_j^k \sim \Gamma(n_z,1)$ [34].

In each cell, the BS follows a UE-specific resource allocation to dedicate a sufficient amount of resources to the served UEs. However, if the SINR of a UE is very small, near infinite bandwidth would be needed to serve the UE. To cope with this, we apply a truncated SINR policy such that the bandwidth allocated to serve UE k is:

$$w_k^{\mathrm{UC}} = \begin{cases} R/\log_2(1+\gamma_k^{\mathrm{UC}}), & \gamma_k^{\mathrm{UC}}\geq\gamma \\ 0, & \text{otherwise} \end{cases}, \quad (8)$$

where γ is the SINR threshold. Hence, no resource is used if the UE is in a poor fading condition.

We assume that a each BS serves all of its U associated UEs by first clustering them into U/n_z groups and applying group-specific ZF beamformings for the n_z UEs of each group. Then, all UEs in a specific group are simultaneously served by allocating a radio resource equaling to the maximum value of resources required by those UEs. As such, the resources consumed in each cell is approximately given by the total resources allocated to UEs in that cell divided by n_z. This becomes precise when $\lambda_{\mathrm{ue}}/\lambda_{\mathrm{bs}}\to\infty$, and is a sufficient approximation when $\lambda_{\mathrm{ue}}\gg\lambda_{\mathrm{bs}}$. We assume that all BSs are active during unicast service, so the average resource consumption is determined by the average consumption of a typical Voronoi cell $\mathcal{V}_0$:

$$W^{\mathrm{UC}}(\cdot) = \mathbb{E}\left\{\sum_{k\in\mathcal{V}_0}\frac{w_k^{\mathrm{UC}}}{n_z}\right\}. \quad (9)$$

Note that the expectation is w.r.t the Poisson processes of UEs (Φ_{ue}) and BSs (Φ_{bs}), as well as effective channel gains based on (7). For the resource consumption of unicast layer we have:

Theorem 1. *Consider an interference-limited frequency reuse 1 cellular network where UEs are served by their nearest BS. BS and UE locations follow PPPs with intensities λ_{bs} and $\lambda_{\mathrm{ue}}^{\mathrm{eff}}$, respectively. BSs use ZF beamforming with L antennas towards n_z simultaneous UEs, and allocate bandwidth to UEs*

using service threshold γ as in (8). The average amount of resources needed is then:

$$W^{\mathrm{UC}}(\gamma,\lambda_{\mathrm{ue}}^{\mathrm{eff}})=\frac{\lambda_{\mathrm{ue}}^{\mathrm{eff}}}{2u\lambda_{\mathrm{bs}}}\sum_{l=1}^{L-n_z+1}f_{l,n_z}\int_0^{w_t}w\frac{d}{dw}C_{l,n_z}(w)dw, \quad (10)$$

where

$$C_{l,n_z}(w)=\frac{1}{{}_2F_1\big(-\frac{2}{e},n_z,1-\frac{2}{e},-\xi\, l\,\eta(w)\big)},$$

$f_{l,n_z}=(-1)^{l+1}\binom{L-n_z+1}{l}$, $w_t=\dfrac{R}{\log_2(1+\gamma)}$, $\xi=\big(L-n_z+1\big)!^{\frac{-1}{L-n_z+1}}$, $\eta(w)=2^{R/w}-1$, *and* ${}_2F_1(\cdot)$ *is the hypergeometric function.*

In addition, The outage probability of a UE is:

$$\mathcal{O}^{\mathrm{UC}}(\gamma)=1-\sum_{l=1}^{L-n_z+1}\frac{f_{l,n_z}}{{}_2F_1(-\frac{2}{e},n_z,1-\frac{2}{e},-\xi\, l\,\gamma)}.$$

Proof. Please refer to the Appendix. □

Note that $W^{\mathrm{UC}}(\cdot)$ depends on $\lambda_{\mathrm{ue}}^{\mathrm{eff}}/\lambda_{\mathrm{bs}}$, not on λ_{bs} and $\lambda_{\mathrm{ue}}^{\mathrm{eff}}$ separately. Note the behavior of $W^{\mathrm{UC}}(\cdot)$ w.r.t $n_z\in\{1,\dots,L\}$; it decreases until reaching a minimum value and then increases, for different values of path-loss exponent e and number of antennas L. Moreover, $\mathcal{O}^{\mathrm{UC}}(\cdot)$ is monotonically increasing w.r.t n_z for fixed L, and monotonically decreasing w.r.t L for fixed n_z. Further, $\mathcal{O}^{\mathrm{UC}}(\cdot)$ is insensitive to the BS intensity λ_{bs}; please refer to Appendix and (32). Finally, $W^{\mathrm{UC}}(\cdot)$ and $\mathcal{O}^{\mathrm{UC}}(\cdot)$ reduces and increases, respectively, as γ_{th} grows.

Corollary 1. *When BSs have a single antenna ($L=1$) and the path-loss exponent is $e=4$, the average resource consumption and outage probability of Theorem 1 become:*

$$W^{\mathrm{UC}}(\gamma,\lambda_{\mathrm{ue}}^{\mathrm{eff}})=\frac{\lambda_{\mathrm{ue}}^{\mathrm{eff}}}{2\lambda_{\mathrm{bs}}}\int_0^{w_t}w\frac{d}{dw}\frac{1}{1+\sqrt{\eta(w)}\tan^{-1}\left(\sqrt{\eta(w)}\right)}dw,$$

$$\mathcal{O}^{\mathrm{UC}}(\gamma)=\frac{\sqrt{\gamma}\tan^{-1}\left(\sqrt{\gamma}\right)}{1+\sqrt{\gamma}\tan^{-1}\left(\sqrt{\gamma}\right)},$$

where $\tan^{-1}(\cdot)$ *is the inverse tangent function.*

Note that Corollary 1 relates to the results given in [35] Eqn. (28), for the case of one-order Voronoi tessellation. Path loss exponents of the order of $e=4$ are typical for moderate and large distances in cellular networks [32].

C. Resource Consumption of Hybrid Scheme

By comparing the results of Theorem 1, we can relate the total resource W^{UC} to the outage probability $\mathcal{O}^{\mathrm{UC}}$:

$$\frac{dW^{\mathrm{UC}}(\gamma,\lambda_{\mathrm{ue}}^{\mathrm{eff}})}{d\gamma}=-\frac{w_t\,\lambda_{\mathrm{ue}}^{\mathrm{eff}}}{2u\,\lambda_{\mathrm{bs}}}\frac{d\mathcal{O}^{\mathrm{UC}}(\gamma)}{d\gamma}. \quad (11)$$

This directly provides a trade-off between the outage probability, and resource consumption of the unicast scheme. As $\dfrac{d\mathcal{O}^{\mathrm{UC}}(\gamma)}{d\gamma}>0$, the increase in w_t makes W^{UC} grows monotonically. The less the service threshold is, the less the service outage is but the more resources are consumed.

Not all UEs request from the unicast component; whereas only UEs being dissatisfied by the multicast component do. As such, based on thinning property of PPP, $\lambda_{\mathrm{ue}}^{\mathrm{eff}}$ in (10) depends on the *overall outage probability of multicast component*, that is obtained by:

$$\mathcal{O}^{\mathrm{MC}}(\boldsymbol{p},\boldsymbol{\alpha})=\sum_{n=1}^{N}f_n\mathcal{O}_n^{\mathrm{MC}}(p_n,\boldsymbol{\alpha}),$$

considering f_n as the popularity of file n, where $\boldsymbol{p}=[p_i]_i$. Therefore, the average of total resource of the network is:

$$W_{\mathrm{tot}}=\underbrace{N_{\mathrm{hb}}R\sum_{n=1}^{N}\beta_n}_{\text{multicast component}}+\underbrace{W^{\mathrm{UC}}(\gamma,\lambda_{\mathrm{ue}})\mathcal{O}^{\mathrm{MC}}(\boldsymbol{p},\boldsymbol{\alpha})}_{\text{unicast component}}. \quad (12)$$

Accordingly, we can obtain the service outage probability $\mathcal{O}_{\mathrm{tot}}$, defined as the probability that a typical UE being served by the hybrid scheme is in outage. We can get:

$$\mathcal{O}_{\mathrm{tot}}=\mathcal{O}^{\mathrm{UC}}(\gamma)\mathcal{O}^{\mathrm{MC}}(\boldsymbol{p},\boldsymbol{\alpha}).$$

IV. Service Latency Analysis

A. Latency in the Multicast Component

For the multicast component applying the HB protocol, the average latency for a typical UE depends on the entire operation of OMPMC. This tier-specific latency is experienced by a typical UE to receive a streamed file from the beginning. According to the HB protocol, if the streamed file is not in multicast outage, the average latency is: $\mathcal{L}^{\mathrm{MC}}=\dfrac{S}{2\,D_{\mathrm{hb}}}$, where S stands for the duration of the broadcasted file in seconds, D_{hb} is the number of HB segments, and the factor $1/2$ is included to account for the temporal statistics of UEs joining the network at different times within a time-slot. However, in the case of multicast outage, the request will be transferred to the nearest BS of the SPUC component. In this case, the UE experiences a BS-specific latency.

B. Latency in the Unicast Component

Within the SPUC layer, there exist multiple requesting users with the average number of $(\lambda_{\mathrm{ue}}/\lambda_{\mathrm{bs}})\mathcal{O}^{\mathrm{MC}}(\boldsymbol{p},\boldsymbol{\alpha})$ per Voronoi cell. Each UE is responded by its nearest BS. Theorem 1 expresses $W^{\mathrm{UC}}(\cdot)$ as the total resource consumption, being averaged over all Voronoi cells, needed to serve requesting UEs. For a typical BS applying $W^{\mathrm{UC}}(\cdot)$ radio resources, there is no unicast latency for its associated UEs if the amount of demanded resources is less than $W^{\mathrm{UC}}(\cdot)$. However, the BS might be requested for more resources than $W^{\mathrm{UC}}(\cdot)$. When that happens, some UEs cannot be served in the time-slot they are requesting service. For this, the BS places those UEs in a queue to serve them during next time-slots. Notice that $W^{\mathrm{UC}}(\cdot)$ can be considered as a baseline value of total required resources, such that the unicast component might reserve slightly more radio resources than $W^{\mathrm{UC}}(\cdot)$ to guarantee the stability of the queue.

More specifically, for a typical Voronoi cell $\mathcal{V}_0$, the associated UEs, need resources $w_k^{\mathrm{UC}}=R/\log_2\left(1+\gamma_k^{\mathrm{UC}}\right),\ k\in\mathcal{V}_0$,

which depends on the statistics of PPPs $\Phi_{\rm ue}$ and $\Phi_{\rm bs}$ and the effective channel coefficients based on (7). The BS responsible of $\mathcal{V}_0$ leverages group-specific ZF beamformings of Section III-B. It then cumulatively sums up the group-wise resource consumptions, computes the difference between the available bandwidth $W^{\rm UC}(\cdot)$ and the cumulative sum, and excludes a group from the sum and places it in the queue if it cannot be served with the current resources. The queued UEs will be then addressed in the next time-slots. Consequently, UEs experience different service latency. We denote the average latency of SPUC by $\mathcal{L}^{\rm UC}$.

V. Optimal Traffic Offloading for Hybrid Scheme

A. Problem Formulation

To design an optimum streaming policy, we minimize the total resource consumption of the hybrid scheme with respect to the cache weights $\boldsymbol{p} = [p_i]_i$, the resource weights $\boldsymbol{\beta} = [\beta_i]_i$, and the multiuser MIMO spatial multiplexing n_z of the cellular base stations. We formulate this problem as a function of service threshold γ:

$$\mathrm{P}_0(\gamma): \min_{\boldsymbol{p},\,\boldsymbol{\beta},\,n_z} \sum_{n=1}^{N}\beta_n + \kappa(\gamma)\sum_{n=1}^{N} f_n \mathcal{O}_n^{\rm MC}\big(p_n, \boldsymbol{\beta}^{-1}\big)$$
$$\text{s.t.} \begin{cases} \beta_n \geq 0, & 0 \leq p_n \leq 1, \quad n \in S_N, \\ \sum_{n=1}^{N} p_n \leq M, & n_z \in \{1,\ldots,L\}, \end{cases}$$

where we define the normalized unicast resource consumption as:

$$\kappa(\gamma) := \frac{W^{\rm UC}(\gamma, \lambda_{\rm ue})}{N_{\rm hb}\, R}. \tag{13}$$

Note that $\mathrm{P}_0(\gamma)$ is parameterized by $\gamma \in \Gamma$, where $\Gamma \subseteq \mathbb{R}^+$ is a set of service thresholds of interest. This is a non-convex mixed-integer optimization problem since n_z belongs to the integer set $\{1,\ldots,L\}$, and $\mathcal{O}_n^{\rm MC}\big(p_n, \boldsymbol{\beta}^{-1}\big)$ is not a convex function based on its Hessian matrix w.r.t. $\boldsymbol{\beta}$ [23]. In the sequel, we characterize the solution of $\mathrm{P}_0(\gamma)$. By defining

$$w_{\rm tot,n} := \beta_n + \kappa(\gamma)\mathcal{O}_n^{\rm MC}\big(p_n, \boldsymbol{\beta}^{-1}\big),$$

some properties are drawn as:

- **Property 1**: If $\beta_n = 0$, then $w_{\rm tot,n} = \kappa(\gamma)$ regardless of the value of p_n. Likewise, if $p_n = 0$, then $w_{\rm tot,n} = \kappa(\gamma)$ from the optimality perspective, regardless of the value of β_n.
- **Property 2**: $w_{\rm tot,n}$ is a monotonically decreasing function w.r.t. β_n and p_n.
- **Property 3**: Without loss of generality, we sort files based on their popularity. Hence, Property 2 dictates that values of p_n and β_n should be decreasing w.r.t. n from the optimality perspective.
- **Property 4**: As a consequence of Property 1, there exists a $K \in S_N$ such that $\beta_n = p_n = 0$ for $n > K$. It reveals that the optimum solution classifies the files into two sets, popular and less popular ones. Additionally, the less popular set is not served by the multicast component, where popular set is served by the multicast component, and then unicasted in the case of multicast outage. We call K the *classifier index*.
- **Property 5**: For given n_z, $\boldsymbol{p}$ and $\boldsymbol{\beta}$, the objective function of $\mathrm{P}_0(\gamma)$ decreases as a function of K till reaching a minimum point and then starts to grow.
- **Property 6**: Based on Property 2, at the optimum solution of $\mathrm{P}_0(\gamma)$, the cache capacity constraint is satisfied with equality; $\sum_{n=1}^{N} p_n = M$.

Proposition 1. *Consider problem $P_0(\gamma)$ for a fixed classifier index K. Solutions $\{p_n^*, \beta_n^*\}_n$ of the Karush– Kuhn–Tucker (KKT) conditions are of two types. Either (i) p_n^* and β_n^* are both decreasing w.r.t. n, or (ii) p_n^* and β_n^* are both increasing w.r.t. n.*

Proof. For simplicity, we show $\kappa(\gamma)$ with κ. Let us define $\sigma_\beta := \sum_{n=1}^{N}\beta_n$, we then use the notation $\mathcal{O}_n^{\rm MC}(p_n, \beta_n^{-1}, \sigma_\beta)$ based on (6) to stress its dependency on p_n, β_n and σ_β. For convenience, we may also use $\mathcal{O}_n^{\rm MC}$.

By formulating the Lagrangian function for $\mathrm{P}_0(\gamma)$ and writing the KKT conditions, we get:

$$\kappa f_n \frac{d\mathcal{O}_n^{\rm MC}}{dp_n} + v_1 = 0, \quad n \in S_K \tag{14a}$$
$$\kappa f_n \frac{d\mathcal{O}_n^{\rm MC}}{d\beta_n} + v_2 = 0, \quad n \in S_K \tag{14b}$$
$$1 + \kappa \sum_{m=1}^{N} f_m \frac{d\mathcal{O}_m^{\rm MC}}{d\sigma_\beta} - v_2 = 0, \tag{14c}$$

where $S_K = \{1,\ldots,K\}$, and v_1 and v_2 are the Lagrangian multipliers related to the equality constraint and definition of σ_β, respectively. On the other hand, based on (6), we have:

$$\frac{d\mathcal{O}_n^{\rm MC}}{d\beta_n} = \frac{2\log(2)}{e}\frac{p_n\, 2^{1/\beta_n}}{\beta_n^2\big(2^{1/\beta_n}-1\big)}\frac{d\mathcal{O}_n^{\rm MC}}{dp_n}, \tag{15a}$$
$$\frac{d\mathcal{O}_n^{\rm MC}}{d\sigma_\beta} = -\frac{2p_n}{e\,\sigma_\beta}\frac{d\mathcal{O}_n^{\rm MC}}{dp_n}. \tag{15b}$$

From (14a), (14b) and (15a), we have

$$p_n = v_e \beta_n^2\left(1 - 2^{-1/\beta_n}\right), \tag{16}$$

where $v_e = \dfrac{v_2}{v_1}\dfrac{e}{2\log(2)}$. According to Property 2 ($\frac{d\mathcal{O}_n^{\rm MC}}{dp_n} \leq 0$) and (14a), we obtain: $v_1 \geq 0$. Also, based on (14b) and (15a), we get: $v_2 \geq 0$, which together yields: $v_e \geq 0$. Further, note that $\beta_n^2\left(1-2^{-1/\beta_n}\right)$ is an increasing function w.r.t. β_n. Based on these and (16), it is concluded that if p_n increases w.r.t. n, β_n should also do. By combining (14a), (14b) and (15b), we obtain $\sigma_\beta = K v_e'$, where $v_e' = \dfrac{2v_1}{e(v_2-1)}$. Consequently, by plugging p_n from (16) into (14a), we can obtain a relationship between f_n and β_n as:

$$\frac{v_1}{\kappa\, f_n} = F_e\big(v_e\beta_n^2\left(1-2^{-1/\beta_n}\right), \beta_n^{-1}, K v_e'\big), \tag{17}$$

where $F_e(p_n, \beta_n^{-1}, \sigma_\beta) = -\dfrac{d\mathcal{O}_n^{\rm MC}(p_n, \beta_n^{-1}, \sigma_\beta)}{dp_n}$. By investigating the curvature of $F_e\big(v_e\beta_n^2\left(1-2^{-1/\beta_n}\right), \beta_n^{-1}, K v_e'\big)$, it is revealed that it grows as a function of β_n till reaching a maximum, and then decreases, for different values of v_e and

v_e'. It shows that solving (17) gives only two solutions, one in the increasing region of $F_e(\cdot)$ and the other in the decreasing region. Considering that f_n is decreasing as a function of n, $F_e\big(v_e\beta_n^2\,(1-2^{-1/\beta_n}),\beta_n^{-1},Kv_e'\big)$ should grows w.r.t. n. This fact leads to two types of KKT solutions for $(\boldsymbol{p},\boldsymbol{\beta})$: (i) The solution that β_n reduces w.r.t. n, and p_n should thus be decreasing w.r.t. n based on (16); (ii) The solution that β_n increases w.r.t. n, and p_n grows w.r.t. n as a consequence of (16). □

Corollary 2. *Consider an approach that solves the KKT conditions of $P_0(\gamma)$, resulting in $\{p_n,\beta_n\}_n$ such that p_n and β_n are both decreasing w.r.t. n. This approach achieves the globally optimal solution.*

This corollary is a consequence of Proposition 1 and Property 3. More specifically, note that the globally optimal solution of any optimization problem is necessarily among all its possible KKT solutions [36]. For $\mathrm{P}_0(\gamma)$ characterized by K, there exist two possible KKT solutions based on Proposition 1, where the one complying with the Property 3 is the globally optimal one.

To solve problem $\mathrm{P}_0(\gamma)$, we resort to a methodology that gives a solution complying with Corollary 2. More specifically, we use a path-following method [37] to sequentially find the optimal solution $\{p_n^*,\beta_n^*\}_{n=1}^N$ of $\mathrm{P}_0(\gamma)$ for $\gamma\in\Gamma$. We show that the solution provided by this methodology satisfies the KKT conditions as well as the solution is decreasing w.r.t. n. Hence, a globally optimal solution can be guaranteed based on Corollary 2. We thus formulate a multi-parametric optimization problem with a θ-parameterized popularity $a_n(\theta)$ with $\sum_{n=1}^N a_n(\theta)=1$ and $a_{n+1}(\theta)\geq a_n(\theta)$ where

$$f_n = a_n(\theta) \qquad \text{for } \theta\in\Theta,$$

and Θ is a set of popularity parameters of interest.

We initially assume that the optimum value n_z^* of this problem is known. In the sequel, we explain how it can be obtained. Accordingly, this multi-parametric problem, with n_z^* being used, is expressed as:

$$\mathrm{P}_1(\gamma,\theta):\min_{\substack{\boldsymbol{p}(\gamma,\,\theta)\\ \boldsymbol{\beta}(\gamma,\,\theta)}}\sum_{n=1}^N\beta_n(\gamma,\theta)+\kappa(\gamma)\sum_{n=1}^N a_n(\theta)\mathcal{O}_n^{\mathrm{MC}}\big(p_n(\gamma,\theta),\boldsymbol{\beta}^{-1}(\gamma,\theta)\big)$$
$$\text{s.t.}\begin{cases}\beta_n(\gamma,\theta)\geq 0, & 0\leq p_n(\gamma,\theta)\leq 1, \quad n\in S_N,\\ \sum_{n=1}^N p_n(\gamma,\theta)\leq M.\end{cases}$$

We additionally consider an initial popularity parameter θ_0, such that $a_n(\theta_0)=\frac{1}{N}$ for $n\in S_N$. For this initial parameter, the optimal solution of $\mathrm{P}_1(\gamma,\theta_0)$ can be simply found as the popularity is uniformly distributed, i.e., $a_n(\theta_0)=\frac{1}{N}$. Then, to obtain the optimal solution $\{p_n^*,\beta_n^*\}_n$ related to $\mathrm{P}_1(\gamma,\theta)$ for $\theta\in\Theta$, we apply an approach to extrapolate the solution from parameter θ_0 into the target popularity parameter θ. We will present this approach in Proposition 2. However, to find n_z^*, we designate three values: the optimum beamforming parameter of $\mathrm{P}_1(\gamma,\theta_0)$, denoted by n_z^0, and unit-length incremented/decremented ones, i.e., n_z^0+1 and n_z^0-1. Then, n_z^* takes one of these three values, based on the numerical evaluations. However, the applied methodology is not restricted to this selection, as a small line search for $n_z\in\{1,\ldots,L\}$ might be instead used.

Notice that the same methodology can also be leveraged to extrapolate the optimal solution from an initial threshold parameter γ_0 to any target parameter $\gamma\in\Gamma$.

We now turn to $\mathrm{P}_1(\gamma,\theta_0)$ with $a_n(\theta_0)=\frac{1}{N}$. By applying Property 4, its optimum cache weights are $p_n^*=\frac{M}{K}$ for $n\in S_K$, and $p_n^*=0$ otherwise, and its optimum resource weights are $\beta_n^*=\beta_0$ for $n\in S_K$, and $\beta_n^*=0$ otherwise. Problem $\mathrm{P}_1(\gamma,\theta_0)$ can be then expressed as:

$$\mathrm{P}_1(\gamma,\theta_0):\min_{\substack{\beta_0>0,\\ n_z\,\in\,\{1,\,\ldots,\,L\}}}\beta_0+\frac{\kappa(\gamma)}{N}\left(\mathcal{O}_n^{\mathrm{MC}}\Big(\frac{M}{K},\frac{1}{\beta_0}\Big)+\frac{K-N}{K}\right).$$

The optimum value of β_0 and n_z is obtained by a gradient descent and a line search, respectively. As declared, it leads to the initial points $\{p_n^*,\beta_n^*\}_n(\gamma,\theta_0)$ and n_z^0 that is used to solve $\mathrm{P}_1(\gamma,\theta)$.

It is noteworthy that for the optimal value of K, we first find a coarse estimation by solving the following optimization problem:

$$\mathrm{P}_2(\gamma,\theta):\min_{\substack{\beta_0>0,\\ K\,\in\,\{M,\,\ldots,\,N\}}}K\beta_0+\kappa(\gamma)\mathcal{O}^{\mathrm{MC}}\Big(\frac{M}{K},\beta_0^{-1}\mathbf{1}\Big)\sum_{n=1}^K a_n(\theta)+\kappa(\gamma)\sum_{n=K+1}^N a_n(\theta),$$

and then search for the fine value by applying Property 5. This considerably reduces the computational complexity.

In the line with explained methodology, we thus have:

Proposition 2. *Consider the multi-parametric optimization problem $P_1(\gamma,\theta)$ characterized by classifier index K. The KKT solution for given $\gamma\in\Gamma$ with dependency on $\theta\in\Theta$ can be found by solving the Ordinary Differential Equation (ODE):*

$$\frac{d\boldsymbol{x}_1(\gamma,\theta)}{d\theta}=-\boldsymbol{A}^{-1}(\gamma,\theta)\,\boldsymbol{b}\big(\dot{\boldsymbol{a}}(\theta)\big), \tag{18}$$

where $\boldsymbol{x}_1(\gamma,\theta)=[\boldsymbol{\beta}^{*\top},\boldsymbol{p}^{*\top},v_3]^\top(\gamma,\theta)$, $v_3=\dfrac{v_1}{\kappa(\gamma)}$ and:

$$\boldsymbol{A}(\gamma,\theta)=\begin{pmatrix}\Big[a_i(\theta)\dfrac{d^2\mathcal{O}_i^{\mathrm{MC}}}{dp_i d\beta_j}\Big]_{i,j} & \Big[a_i(\theta)\dfrac{d^2\mathcal{O}_i^{\mathrm{MC}}}{dp_i^2}\delta_{ij}\Big]_{i,j} & \mathbf{1}\\ \Big[\sum_{n=1}^K a_n(\theta)\dfrac{d^2\mathcal{O}_n^{\mathrm{MC}}}{d\beta_i d\beta_j}\Big]_{i,j} & \Big[a_j(\theta)\dfrac{d^2\mathcal{O}_j^{\mathrm{MC}}}{dp_j d\beta_i}\Big]_{i,j} & \mathbf{0}\\ \mathbf{0}^\top & \mathbf{1}^\top & 0\end{pmatrix}^{-1}, \tag{19}$$

where $i,j\in S_K$ and for vector $\boldsymbol{c}=[c_1,\ldots,c_K]^\top$, we have:

$$\boldsymbol{b}\big(\boldsymbol{c}\big)=\left(\Big[c_i\frac{d\mathcal{O}_i^{\mathrm{MC}}}{dp_i}\Big]_i^\top,\quad\Big[\sum_{n=1}^K c_n\frac{d\mathcal{O}_n^{\mathrm{MC}}}{d\beta_i}\Big]_i^\top,0\right)^\top, \tag{20}$$

where $i\in S_K$.
Moreover, the KKT solution of $P_1(\gamma,\theta)$ for given $\theta\in\Theta$ with dependency on $\gamma\in\Gamma$, can be found by solving the ODE:

$$\frac{d\boldsymbol{x}_2(\gamma,\theta)}{d\gamma}=-\frac{d\kappa(\gamma)}{d\gamma}\frac{1}{\kappa(\gamma)}\boldsymbol{A}^{-1}(\gamma,\theta)\,\boldsymbol{b}\big(\boldsymbol{a}(\theta)\big), \tag{21}$$

$$\boldsymbol{d} = \left(\left[a_i(\theta) \frac{d^2 \mathcal{O}_i^{\mathrm{MC}}}{dp_i^2} \dot{p}_i + \sum_{m=1}^{K} a_i(\theta) \frac{d^2 \mathcal{O}_i^{\mathrm{MC}}}{dp_i d\beta_m} \dot{\beta}_m \right]_i^{\top}, \left[\sum_{n=1}^{K} a_n(\theta) \frac{d^2 \mathcal{O}_n^{\mathrm{MC}}}{dp_n d\beta_i} \dot{p}_n + \sum_{m,n}^{K} a_n(\theta) \frac{d^2 \mathcal{O}_n^{\mathrm{MC}}}{d\beta_i d\beta_m} \dot{\beta}_m \right]_i^{\top}, \quad 0 \right)^{\top}. \tag{24}$$

where $\boldsymbol{x}_2(\gamma,\theta) = [\boldsymbol{\beta}^{*\top}, \boldsymbol{p}^{*\top}, v_1]^{\top}(\gamma,\theta)$.

Proof. It can be proved by constituting the Lagrangian function for $\mathrm{P}_1(\gamma,\theta)$, corresponding KKT conditions and applying the Homotopy Continuation [37] at $\theta + d\theta$. □

By solving ODE (18), we can find the optimal solution of $\mathrm{P}_1(\gamma,\theta)$ for different values of $\theta \in \Theta$ and given $\gamma \in \Gamma$.

Note that for the initialization, we use the optimal solution of $\mathrm{P}_1(\gamma,\theta_0)$. This extrapolates the KKT solution from θ_0 into any parameter $\theta \in \Theta$, and consequently provides the basis for the path-following approach we mentioned earlier. Likewise, by solving (21), the optimal solution can be found for given θ and different values of γ.

However, ODE (18) is intricate to be analytically solved due to its non-linearity w.r.t. θ. Instead, we use the Euler method, with an incremental-step $\Delta\theta \in (0,1]$, to sequentially find $\boldsymbol{x}_1(\gamma,\theta)$ [25]. Based on (18), we get:

$$\boldsymbol{x}_1(\gamma,\theta) = \boldsymbol{x}_1(\gamma,\theta-\Delta\theta) + \frac{d\boldsymbol{x}_1(\gamma,\theta-\Delta\theta)}{d\theta}\Delta\theta, \qquad \text{for } \theta \in \Theta.$$

We thus start from the initial points $\boldsymbol{x}_1(\gamma,\theta_0)$, and increment θ by $\Delta\theta$ to obtain the solution $\boldsymbol{x}_1(\gamma,\theta_0+\Delta\theta)$. We repeat this until the optimal solutions for all $\theta \in \Theta$ are found. Note that the Euler method with an incremental-step $\Delta\gamma \in (0,1]$, can also be used to solve ODE (21).

Consequently, the solution of these two ODEs, gives a surface $\boldsymbol{y}(\gamma,\theta) := [\boldsymbol{\beta}^{*\top}, \boldsymbol{p}^{*\top}]^{\top}(\gamma,\theta)$ that portrays the optimal solutions for different file popularities and service thresholds. Hereafter, we name it as the Optimal Surface Solution (OSS). We can relate these two ODEs, in order to obtain the OSS using only one ODE.

Proposition 3. *The optimal surface solution* $\boldsymbol{y}(\gamma,\theta)$ *can be found by solving the second-order ODE:*

$$\frac{d^2\boldsymbol{y}(\gamma,\theta)}{d\theta d\gamma} = -\mathcal{E}\big(\boldsymbol{A}(\gamma,\theta)^{-1}\,\dot{\boldsymbol{A}}(\gamma,\theta)\big)\,\frac{d\boldsymbol{y}(\gamma,\theta)}{d\gamma}, \tag{22}$$

where $\mathcal{E}(\cdot)$ *is a matrix operator that omits the last column and the last row of its argument.*

Proof. From (21) in Proposition 2 we have:

$$\frac{d^2\,\boldsymbol{x}_2}{d\gamma d\theta} = \frac{d\kappa(\gamma)}{d\gamma}\frac{1}{\kappa(\gamma)}\boldsymbol{A}^{-1}\left(\dot{\boldsymbol{A}}\boldsymbol{A}^{-1}\,\boldsymbol{b}\big(\boldsymbol{a}(\theta)\big) - \boldsymbol{b}\big(\dot{\boldsymbol{a}}(\theta)\big) - \boldsymbol{d}\right), \tag{23}$$

where $\boldsymbol{d}$ is defined in (24) on the top of the page. According to (19) and $\sum_{m=1}^{K}\dot{p}_m = 0$, we obtain: $\boldsymbol{A}^{-1}\boldsymbol{d} = \big[\dot{\boldsymbol{\beta}}^{\top}, \dot{\boldsymbol{p}}^{\top}, 0\big]^{\top}$. By plugging this into (23) and using Proposition 2 we get:

$$\frac{d^2\,\boldsymbol{x}_2}{d\gamma d\theta} = -\boldsymbol{A}^{-1}\dot{\boldsymbol{A}}\,\frac{d\,\boldsymbol{x}_2}{d\gamma} + \frac{d\kappa(\gamma)}{d\gamma}\,\frac{1}{\kappa(\gamma)}\,\big(\mathbf{0}^{\top}, \mathbf{0}^{\top}, v_3\big)^{\top}. \tag{25}$$

Considering that the last column of $\boldsymbol{A}^{-1}\dot{\boldsymbol{A}}$ is zero, (22) follows. □

Note that (22) can be interpreted as a sequential approach to extrapolate a γ-parameterized path along θ, which forms a surface w.r.t. γ and θ. Since this approach is on the basis of a path-following method [37], we name it as the surface-following method. Therefore, the solution of (22) for $\theta \in \Theta$ and $\gamma \in \Gamma$ gives the OSS $\{p_n^*, \beta_n^*\}_n(\gamma,\theta)$ that satisfies (14a)-(14c).

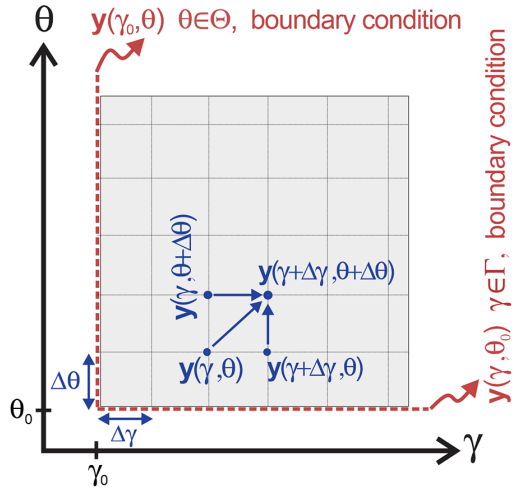

Fig. 4: Numerical solution of second-order ODE (22).

To solve the ODE (22), we need two boundary conditions $\boldsymbol{y}(\gamma,\theta_0)$ for $\gamma \in \Gamma$ and $\boldsymbol{y}(\gamma_0,\theta)$ for $\theta \in \Theta$, which are associated with problems $\mathrm{P}_1(\gamma,\theta_0)$ and $\mathrm{P}_1(\gamma_0,\theta)$, respectively. As declared, θ_0 and γ_0 are the initial parameters of popularity and service threshold, respectively. Moreover, we exploit the Euler method to obtain the solution of (22), assuming K. As such, we consider the incremental-steps $\Delta\theta$ and $\Delta\gamma$, and from (22) we obtain:

$$\begin{aligned}\boldsymbol{y}(\gamma+\Delta\gamma,\theta+\Delta\theta) &= \boldsymbol{y}(\gamma,\theta+\Delta\theta)\\ &\quad - \boldsymbol{A}(\gamma,\theta)^{-1}\dot{\boldsymbol{A}}(\gamma,\theta+\Delta\theta)\Big(\boldsymbol{y}(\gamma+\Delta\gamma,\theta) - \boldsymbol{y}(\gamma,\theta)\Big).\end{aligned} \tag{26}$$

We thus start from the boundary points $\boldsymbol{y}(\gamma_0,\theta_0)$, $\boldsymbol{y}(\gamma_0+\Delta\gamma,\theta_0)$ and $\boldsymbol{y}(\gamma_0,\theta_0+\Delta\theta)$, and increment γ by $\Delta\gamma$ to obtain the solution $\boldsymbol{y}(\gamma_0+\Delta\gamma,\theta_0+\Delta\theta)$ using (26). In the next iteration, we start from $\boldsymbol{y}(\gamma_0+\Delta\gamma,\theta_0)$, $\boldsymbol{y}(\gamma_0+2\Delta\gamma,\theta_0)$ and the recently obtained $\boldsymbol{y}(\gamma_0+\Delta\gamma,\theta_0+\Delta\theta)$, to find $\boldsymbol{y}(\gamma_0+2\Delta\gamma,\theta_0+\Delta\theta)$. We repeat this sequential process over γ till the optimal solutions for all $\gamma \in \Gamma$ are obtained. This finally gives the path solution $\boldsymbol{y}(\cdot,\theta_0+\Delta\theta)$. We now increment θ by $\Delta\theta$ and execute the full incremental process over γ to obtain the path solution $\boldsymbol{y}(\cdot,\theta_0+2\Delta\theta)$. Figure 4 depicts this sequential process for a particular γ and θ. This process is continued till the optimal solutions for all $\theta \in \Theta$ are obtained, which gives the OSS $\boldsymbol{y}(\gamma,\theta)$.

As declared, we need to solve $\mathrm{P}_1(\gamma,\theta_0)$ and $\mathrm{P}_1(\gamma_0,\theta)$ to obtain the boundary conditions for solution of (26). We earlier explained how to solve $\mathrm{P}_1(\gamma,\theta_0)$. However, to solve $\mathrm{P}_1(\gamma_0,\theta)$, we consider the ODE (18) and exploit the solution

of $\mathrm{P}_1(\gamma_0, \theta_0)$ to obtain the initial point. We then apply the same numerical approach as utilized in (26).

Based on the numerical evaluations, the solution of $\mathrm{P}_1(\gamma_0, \theta)$ is decreasing w.r.t. n for any $\theta \in \Theta$. Therefore, based on Corollary 2, the globally optimal solutions for different file popularities are obtained. By having such solution, the obtained OSS $\{p_n^*, \beta_n^*\}_n(\gamma, \theta)$ is also decreasing w.r.t. n for any $\theta \in \Theta$ and $\gamma \in \Gamma$. It shows that, the globally optimal solutions of $\mathrm{P}_0(\gamma)$ is achieved for different file popularities, according to Corollary 2.

B. Solution Analysis

Here, we analyze the developed path-following method and compare it with parallel Successive- Convex-Approximation (p-SCA) [38] as well as with a *benchmark* approach, being obtained with an extremely small incremental step $\Delta\theta = 10^{-6}$ regardless of its computational complexity. The comparison is done from the optimality, convergence and computational complexity perspectives.

1) Convergence and Optimality of Solutions: According to Corollary 2, the path-following solutions provided by Propositions 2 and 3 converge to the global solution. However, we utilize the Euler method with incremental-steps $\Delta\theta$ and $\Delta\gamma$ to approximate the solution. As such, the convergence and optimality of solution depends on the values of the incremental-step. For the benchmark approach with an extremely small value of the incremental-step, the solution converges to the global solution, based on the Corollary 2, at the cost of a computational complexity. However, for the practical values of the incremental-step, a performance gap between the solution and the global optimum exists. To measure it, we do as follows:

- We obtain a benchmark solution with the extremely small incremental-step $\Delta\theta = 10^{-6}$. For the target popularity θ and service threshold γ, we denote the benchmark solution by $(\boldsymbol{p}^*, \boldsymbol{\beta}^*)(\gamma, \theta)$.
- We consider a metric to compare a sub-optimum solution $(\boldsymbol{p}, \boldsymbol{\beta})(\gamma, \theta)$, (either the path-following solution by the Euler method or p-SCA solution), with the globally optimum of Benchmark solution, from the optimality perspective. We then introduce an *optimality distance* as follows:

$$O_d = \left(\|\boldsymbol{p}(\gamma, \theta) - \boldsymbol{p}^*(\gamma, \theta)\|^2 + \|\boldsymbol{\beta}(\gamma, \theta) - \boldsymbol{\beta}^*(\gamma, \theta)\|^2 \right)^{1/2}. \quad (27)$$

It is noteworthy that the convergence of p-SCA depends on a *termination criterion*, the *step-size* of gradient descent as well as the choice of a surrogate function.

2) Computational Complexity of Solutions: For the path-following method, the computational complexity mainly depends on the selected incremental-step and process of finding the optimum ZF beamforming parameter n_z^*. We choose the incremental-step $\Delta\theta = 10^{-2}$, as it provides a solution with acceptable optimality distance and computational complexity.

For the p-SCA approach, the complexity mainly depends on the step-size, hyper-parameters of the surrogate function, and the number of iterations. We tune the step-size to be 0.1 and set the number of iterations around to 50, as this setting approximately gives the minimum value of optimality distance with low complexity. To compare the path-following method with p-SCA, from the complexity viewpoint, we measure the execution time T_{ex} required for these algorithms to find a solution.

Note that for p-SCA and path-following methods, we use the solution of $\mathrm{P}_1(\gamma, \theta_0)$ for the algorithm initialization. For both, we include the requirement time of $\mathrm{P}_1(\gamma, \theta_0)$ in T_{ex}.

VI. Simulation Results and Discussion

We apply the developed Path-Following (PF), p-SCA [38] and benchmark (Bench) approaches to find the optimum solution of hybrid cache policy. We then compare the optimal solution to conventional single-antenna SPUC [12], [39], [40], multi-antenna ZF beamforming SPUC [4], [34], [41] and OMPMC schemes [23] from the literature.

We consider the following scenario. The number of files is $N = 100$, the cache capacity of HNs $M = 10$ and the streaming service rate $R = 1$ Mbps. For the file popularity, we use the Zipf distribution [42] with skewness θ: $f_n = n^{-\theta} / \sum_{m=1}^{N} m^{-\theta}$. As such, the skewness θ is parameterizing the file popularity. The skewness of interest is set to $\theta = \{0.6, 2.0\}$, and the initialization point is $\theta_0 = 0$, as $f_n = \frac{1}{N}$ for $\theta = 0$. We have $L = 8$ antennas at the BSs and the service threshold of interest $\gamma = 0.1$, unless they are specified. We set the Harmonic number $N_{\mathrm{hb}} = 6$, which increases the effective bandwidth of OMPMC component with a factor of 6, but remarkably reduces the streaming latency with a factor of $D_{\mathrm{hb}} = 226$ as $\sum_{i=1}^{226} \frac{1}{i} \approx 6$ [28]. As an example, the average multicast latency of a file with duration of one hour becomes $\mathcal{L}^{\mathrm{MC}} = \frac{1}{2} \frac{3600}{226} \approx 8$ seconds. The BSs and HNs are deployed based on two PPPs with intensity $\lambda_{\mathrm{bs}} = 200$ and $\lambda_{\mathrm{hn}} \in [5, 200]$, respectively, and UEs are located according to another PPP with intensity $\lambda_{\mathrm{ue}} = 2 \times 10^5$. We apply an Urban NLOS scenario from 3GPP [43] with carrier frequency 2 GHz, HN transmission power 23 dBm, and path-loss exponent $e = 3.76$. As $\lambda_{\mathrm{ue}} \gg \lambda_{\mathrm{bs}}$, the network of BSs operates in an interference-limited regime. The antenna gains at the UE and HN are 0 dBi and 8 dBi, the noise-figure of UE is 9 dB, the noise spectrum density is -174 dBm. The reference distance is 1 km, so the PPP intensities are in the units of points/km^2.

Figure 5 shows the normalized resource usage for the hybrid scheme, and the SPUC and OMPMC components as a function of HN intensity λ_{hn}, being obtained by Path-Following (PF), benchmark (Bench) and p-SCA approaches. The optimum resource consumption of PF and the benchmark methods are close to each other, but p-SCA solution slightly differs, which shows the precision of PF solution w.r.t the global optimum. As λ_{hn} increases, the total resource consumption decreases– the OMPMC component consumes more resources, as more files are offloaded to be served in OMPMC.

Figure 6 illustrates the outage probability as a function of HN intensity λ_{hn}, being obtained by different approaches. The optimum values of outage probability of PF and the benchmark methods matche, but the p-SCA solution slightly differs. Note that the outage probability of the SPUC component is insensitive to λ_{hn}. As λ_{hn} increases, the total outage probability

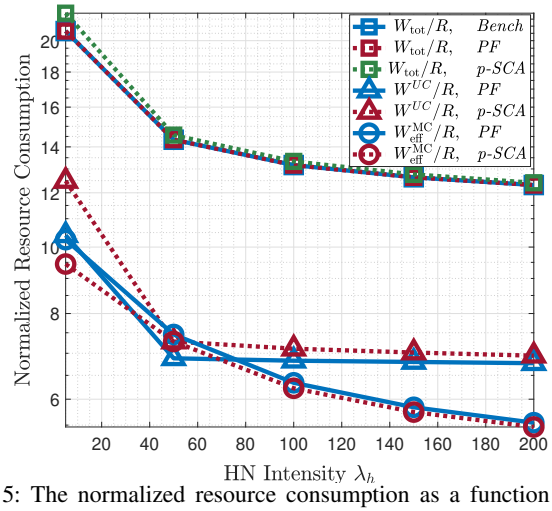

Fig. 5: The normalized resource consumption as a function of HN intensity λ_{hn} for $\lambda_{\text{bs}} = 200$, $\theta = 2.0$.

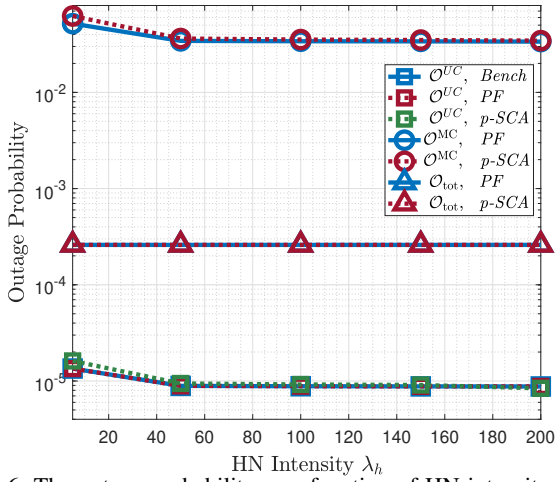

Fig. 6: The outage probability as a function of HN intensity λ_{hn} for $\lambda_{\text{bs}} = 200$, $\theta = 2.0$.

of the hybrid scheme decreases due to the increase in the performance of the OMPMC component.

We now compare path-following and p-SCA solutions w.r.t the benchmark one, from the optimality perspective using metric (27), and the execution time T_{ex} as a metric of computational complexity. We use the settings of Figure 5 for the comparison. Table I compares these approaches from the optimality and complexity viewpoints as a function of λ_{hn} for $\lambda_{\text{bs}} = 200$. It can be seen that the complexity of PF is less than p-SCA, and it can considerably outperform p-SCA from the optimality distance perspective. This confirms that PF can approximate the globally optimal solution with an acceptable precision. As such, hereafter, we focus on PF approach for other simulation setups.

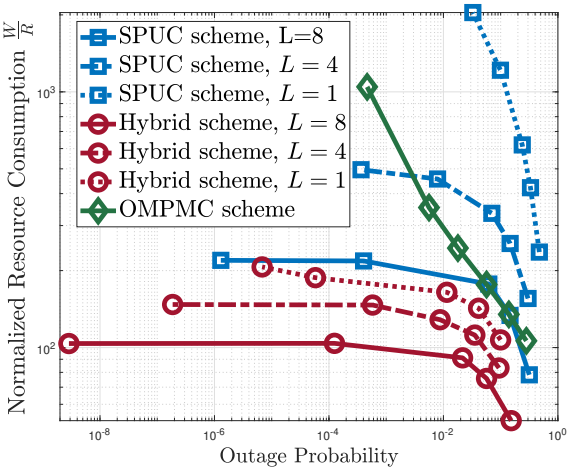

Fig. 7: The normalized resource consumption versus outage probability for $\lambda_{\text{bs}} = 200$, $\lambda_{\text{hn}} = 50$, $\theta = 0.6$.

To investigate how the proposed hybrid scheme optimizes the total resource usage of the network, we compare it with OMPMC scheme, as well as SPUC transmission with the single-antenna and multi-antenna beamforming. Figure 7 portrays the tradeoff between total resource usage and outage probability of the different schemes for $\lambda_{\text{bs}} = 200$, and a sparse MPMC component with $\lambda_{\text{hn}} = 50$. To generate the curves of SPUC and hybrid schemes, we change the service threshold in the range $\gamma \in [0.03, 1]$. For the OMPMC scheme, we target a value for the outage probability and determine the corresponding amount of resource consumption. The OMPMC scheme performs worst as compared to the single-antenna SPUC and the hybrid schemes. The reason for this is that the considered Zipf distribution with skewness $\theta = 0.6$ has a fat tail—there is a significant number of requests for unlikely cached files. Despite this, the hybrid scheme is able to considerably outperform the SPUC scheme for different values of L, offloading the most popular files to the OMPMC component. Moreover, as the number of antennas grows, the performance of hybrid scheme improves.

The same tradeoff for a cellular network, where $\lambda_{\text{bs}} = \lambda_{\text{hn}} = 200$, is depicted in Figure 8. In this case, the BS and caching HNs can be considered to be the same. Now, OMPMC generically outperforms SPUC for any number of antennas. The reason is that densifying the HNs increases the possibility of caching low popular files. However, the hybrid scheme still provides the best service with a wide margin.

Figure 9 shows the spectral efficiency of traffic streaming policy as a function of the file index for $\lambda_{\text{bs}} = 200$ and service threshold $\gamma = 0.1$. According to Property 4, the values not being sketched in this figure have zero values of resource weights (β_n) or infinite spectral efficiency. The solutions are increasing w.r.t. n for all evaluated intensity λ_{hn}, as declared

TABLE I: Performance comparison between p-SCA and Path-Following (PF) approaches.

Approach	Execution time T_{ex} [s]				Optimality distance O_d			
	$\lambda_{\text{HN}} = 5$	$\lambda_{\text{HN}} = 50$	$\lambda_{\text{HN}} = 100$	$\lambda_{\text{HN}} = 200$	$\lambda_{\text{HN}} = 5$	$\lambda_{\text{HN}} = 50$	$\lambda_{\text{HN}} = 100$	$\lambda_{\text{HN}} = 200$
PF	13.3	14.0	13.3	9.6	7.4×10^{-4}	3.0×10^{-4}	2.5×10^{-4}	2.16×10^{-4}
p-SCA	18.0	22.5	19.8	20.1	0.11	0.056	0.055	0.050

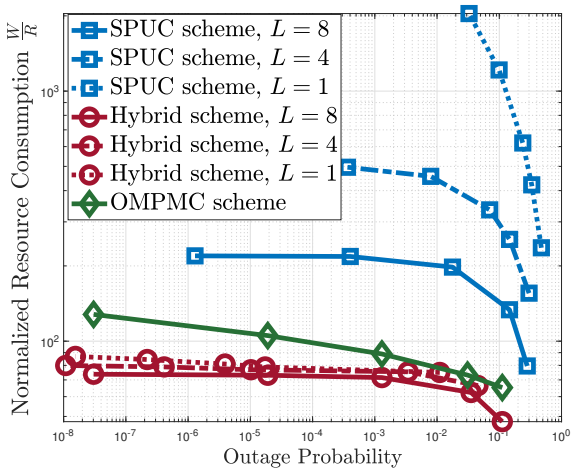

Fig. 8: The Normalized resource versus outage probability for $\lambda_{\rm bs} = \lambda_{\rm hn} = 200$, $\theta = 0.6$.

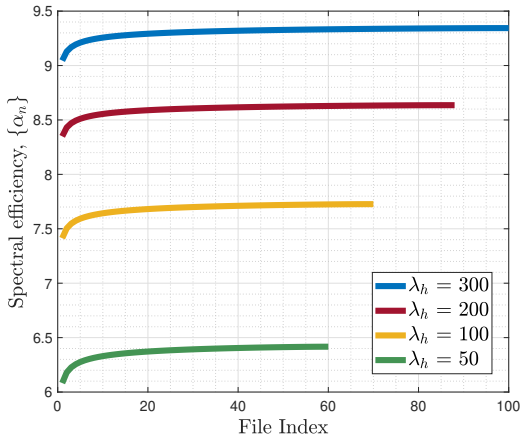

Fig. 9: Bandwidth allocation as a function of file index for $\lambda_{\rm bs} = 200$ and $\gamma = 0.1$.

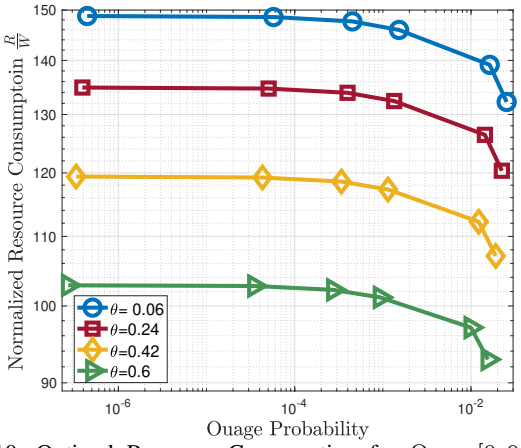

Fig. 10: Optimal Resource Consumption for $\Theta = [0, 0.6]$, $\Gamma = [0.03, 0.3]$ and for $\lambda_{\rm hn} = 50$, $\lambda_{\rm bs} = 200$.

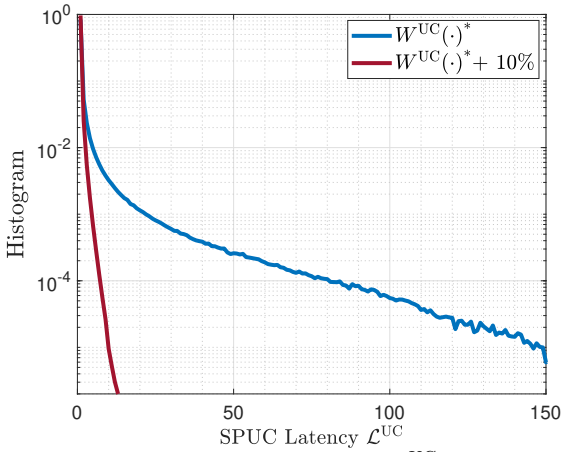

Fig. 11: The histogram of unicast latency $\mathcal{L}^{\rm UC}$ in the unit of slots with duration of multicast latency $\mathcal{L}^{\rm MC}$.

in Property 3. According to Corollary 2, it shows that the used solution approach has been able to find the global optimum. As HN intensity increases, the classifier index K and spectral efficiency α_n grow, which shows a trade-off between K and OMPMC file-specific resource $w_n^{\rm MC}$. This shows that as HN intensity increases, more files are offloaded to the OMPMC component for streaming.

Figure 10 shows the optimal value of resource consumption for sets $\Theta = [0, 0.6]$ and $\Gamma = [0.03, 0.3]$ and for $\lambda_{\rm hn} = 50$, $\lambda_{\rm bs} = 200$. The OSS has been computed based on Proposition 3 and (26), by which the amount of total resource consumption and service outage probability has been obtained. As the skewness decreases which lead to a more flattened file popularity, the amount of total resource consumption grows for the fixed service outage probability. Moreover, the service outage probability improves as the skewness increases for given total resource consumption.

We now investigate the unicast latency $\mathcal{L}^{\rm UC}$ based on the exposition of Section IV-B. For this, we set the intensities of UEs, BSs and HNs to $\frac{\lambda_{\rm ue}}{\lambda_{\rm bs}} = 10^3$ and $\lambda_{\rm hn} = 100$, the number of antennas $L = 8$, the popularity parameter $\theta = 0.6$, the service threshold $\gamma = 0.1$ and the harmonic number $N_{hb} = 6$. Then, we apply the optimum streaming policy being found by path-following method which is characterized by $n_z^* = 3$, $W^{\rm UC}(\cdot)^* = 32$ MHz, $O^{\rm MC}(\cdot)^* = 0.14$, We then perform a numerical simulation to evaluate unicast latency. In addition, we consider the case where the network applies more bandwidth than $W^{\rm UC}(\cdot)^*$ in order to guarantee the stability of UE queue and reducing the unicast latency. Figure 11 depicts the histogram of unicast latency $\mathcal{L}^{\rm UC}$, in the unit of slots with duration of multicast latency $\mathcal{L}^{\rm MC}$, for the cases (i) $W^{\rm UC}(\cdot)^*$ is used, and (ii) $W^{\rm UC}(\cdot)^*$ with an additional 10% is utilized. As this figure shows, for the former case, UEs experience a SPUC latency of $\mathcal{L}^{\rm UC} = 5\mathcal{L}^{\rm MC} = 2.5\frac{S}{D}$ with a probability less than 10^{-2}. For a stream with duration of $S = 1$ hours, this latency thus becomes 40 seconds. For the case with an additional 10% of resources amount, the tail of histogram significantly vanishes faster compared to the previous case. Moreover, UEs experience a latency of $\mathcal{L}^{\rm UC} = 2.5\mathcal{L}^{\rm MC} = 20$ seconds with probability 10^{-2}, which is half of the previous case.

VII. Conclusion

In this paper, we considered a hybrid content streaming system combining orthogonal multipoint multicast (OMPMC) and single-point unicast (SPUC) transmission schemes. We

designed a streaming policy where the optimization is with respect to the probabilistic cache placement policy and the radio resource allocation of OMPMC and multiuser beamforming parameter of SPUC, for given target outage probability. Using the stochastic geometry, we find the followings. The amount of consumed resources depends on the ratio of user intensity to SPUC BS intensity, and not on both separately. Further, a functional relationship between the total resource consumption and the outage probability of unicast service was found. To find the optimal solution, a parametric optimization problem was solved using a path-following method. The solution shows that the files are classified into two sets, popular and less popular ones. OMPMC takes care of serving (most of) the requests to popular files, unless some user is in very bad signal reception conditions. The requests to not-popular content are served by the SPUC component.

We compared the performance of the hybrid scheme with content delivery based on conventional SPUC-only, or on OMPMC-only. Simulation results showed that the hybrid scheme outperforms SPUC with a wide margin. This means that offloading the popular content to OMPMC represents a very attractive option. Moreover, this indicates that eMBMS, can be used in combination with harmonic broadcasting and a suitable caching policy for popular content to achieve significant transmission resource savings in comparison to SPUC-only systems. Hence, the proposed hybrid delivery scheme is a promising candidate for optimizing the spectral efficiency and total resource consumption in heterogeneous and cellular networks.

REFERENCES

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

[2] D. Liu, B. Chen, C. Yang, and A. F. Molisch, "Caching at the wireless edge: design aspects, challenges, and future directions," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 22–28, Sep. 2016.

[3] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2015, pp. 3358–3363.

[4] J. Wu, B. Chen, C. Yang, and Q. Li, "Caching and bandwidth allocation policy optimization in heterogeneous networks," in *Proc. IEEE PIMRC*, Oct. 2017, pp. 1–6.

[5] Y. Chen, M. Ding, J. Li, Z. Lin, G. Mao, and L. Hanzo, "Probabilistic small-cell caching: Performance analysis and optimization," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 4341–4354, 2017.

[6] F. Cheng, G. Gui, N. Zhao, Y. Chen, J. Tang, and H. Sari, "UAV-relaying-assisted secure transmission with caching," *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3140–3153, 2019.

[7] N. Zhao, F. Cheng, F. R. Yu, J. Tang, Y. Chen, G. Gui, and H. Sari, "Caching UAV assisted secure transmission in hyper-dense networks based on interference alignment," *IEEE Trans. Commun.*, vol. 66, no. 5, pp. 2281–2294, 2018.

[8] F. Zhou, L. Fan, N. Wang, G. Luo, J. Tang, and W. Chen, "A cache-aided communication scheme for downlink coordinated multipoint transmission," *IEEE Access*, vol. 6, pp. 1416–1427, Dec. 2018.

[9] J. Laiho, A. Wacker, and T. Novosad, *Radio Network Planning and Optimisation for UMTS*. John Wiley & Sons, 2006.

[10] L. Garcia, K. Pedersen, and P. Mogensen, "Autonomous component carrier selection: Interference management in local area environments for lte-advanced," *IEEE Commun. Mag.*, vol. 47, no. 9, pp. 110–116, 2009.

[11] F. Rusek, D. Persson, B. K. Lau, E. G. Larsson, T. L. Marzetta, O. Edfors, and F. Tufvesson, "Scaling up mimo–opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, 2013.

[12] J. Wu, C. Yang, and B. Chen, "Proactive caching and bandwidth allocation in heterogenous networks by learning from historical numbers of requests," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4394–4410, 2020.

[13] M. Choi, A. F. Molisch, D. Han, D. Kim, J. Kim, and J. Moon, "Probabilistic caching and dynamic delivery policies for categorized contents and consecutive user demands," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2685–2699, 2021.

[14] Y. Li, H. Ma, L. Wang, S. Mao, and G. Wang, "Optimized content caching and user association for edge computing in densely deployed heterogeneous networks," *IEEE Transactions on Mobile Computing*, vol. 21, no. 6, pp. 2130–2142, 2022.

[15] M. Esmaeili, S. Shahbazpanahi, and M. Dong, "Joint optimization of transmit beamforming and base station cache allocation in multi-cell C-RAN," *IEEE Trans. Signal Process.*, vol. 71, pp. 1755–1769, 2023.

[16] Y. Cui and D. Jiang, "Analysis and optimization of caching and multicasting in large-scale cache-enabled heterogeneous wireless networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 1, pp. 250–264, Jan. 2017.

[17] S. Zhong and X. Wang, "Joint multicast and unicast beamforming for coded caching," *IEEE Trans. on Commun.*, vol. 66, no. 8, pp. 3354–3367, 2018.

[18] L. Zhong, C. Xu, J. Chen, W. Yan, S. Yang, and G.-M. Muntean, "Joint optimal multicast scheduling and caching for improved performance and energy saving in wireless heterogeneous networks," *IEEE Trans. Broadcast.*, vol. 67, no. 1, pp. 119–130, 2021.

[19] L. Liu, C. Hua, J. Yu, and J. Xu, "Joint beamforming for delay optimal transmission in cache-enabled wireless backhaul networks," *Journal Commun. and Informat. Networks*, vol. 8, no. 2, pp. 141–154, 2023.

[20] H. Sari, G. Karam, and I. Jeanclaude, "Transmission techniques for digital terrestrial TV broadcasting," *IEEE Commun. Mag.*, vol. 33, no. 2, pp. 100–109, Feb. 1995.

[21] F. X. A. Wibowo, A. A. P. Bangun, A. Kurniawan, and Hendrawan, "Multimedia broadcast multicast service over single frequency network (MBSFN) in LTE based femtocell," in *Proc. Int. Conf. Elect. Eng. Informat.*, Jul. 2011, pp. 1–5.

[22] L. Militano, M. Condoluci, G. Araniti, A. Molinaro, A. Iera, and G. Muntean, "Single frequency-based device-to-device-enhanced video delivery for evolved multimedia broadcast and multicast services," *IEEE Trans. Broadcast.*, vol. 61, no. 2, pp. 263–278, Jun. 2015.

[23] M. Amidzadeh, H. Al-Tous, G. Caire, and O. Tirkkonen, "Caching in cellular networks based on multipoint multicast transmissions," *IEEE Trans. Wireless Commun.*, pp. 1–19, 2022.

[24] M. Amidzadeh, O. Tirkkonen, and G. Caire, "Optimal bandwidth allocation for multicast-cache-aided on-demand streaming in wireless networks," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, 2022, pp. 1936–1941.

[25] M. Amidzade, H. Al-Tous, O. Tirkkonen, and G. Caire, "Orthogonal multipoint multicast caching in OFDM cellular networks with ICI and IBI," in *Proc. IEEE Annu. Int. Symp. Pers. Indoor, Mobile Radio Commun.*, 2021, pp. 394–399.

[26] ——, "Cellular traffic offloading with optimized compound single-point unicast and cache-based multipoint multicast," in *Proc. IEEE Wireless Commun. and Netw. Conf. (WCNC)*, Dec. 2022, pp. 1–6.

[27] S. He, H. Tian, X. Lyu, G. Nie, and S. Fan, "Distributed cache placement and user association in multicast-aided heterogeneous networks," *IEEE Access*, vol. 5, pp. 25 365–25 376, 2017.

[28] L.-S. Juhn and L.-M. Tseng, "Harmonic broadcasting for video-on-demand service," *IEEE Trans. Broadcast.*, vol. 43, no. 3, pp. 268–271, 1997.

[29] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *Proc. IEEE Int. Conf. Comput. Commun., INFOCOM*, 1999, pp. 126–134.

[30] B. Serbetci and J. Goseling, "On optimal geographical caching in heterogeneous cellular networks," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, March. 2017, pp. 1–6.

[31] F. Baccelli and B. Blaszczyszyn, "Stochastic geometry and wireless networks, volume 1: Theory," *Found. Trends Netw.*, vol. 3, no. 3-4, pp. 249–449, 2009.

[32] A. Goldsmith, *Wireless Communications*. Cambridge University Press, 2005.

[33] N. Jindal, J. G. Andrews, and S. Weber, "Multi-antenna communication in ad hoc networks: Achieving MIMO gains with SIMO transmission," *IEEE Trans. Commun.*, vol. 59, no. 2, pp. 529–540, 2011.

[34] Z. Chen, L. Qiu, and X. Liang, "Area spectral efficiency analysis and energy consumption minimization in multiantenna poisson distributed networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4862–4874, 2016.

[35] N. Lee, D. Morales-Jimenez, A. Lozano, and R. W. Heath, "Spectral efficiency of dynamic coordinated beamforming: A stochastic geometry approach," *IEEE Trans. Wireless Commun.*, vol. 14, no. 1, pp. 230–241, 2015.

[36] S. P. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge Univ. Pr., 2011.

[37] V. Kungurtsev and M. Diehl, "Sequential quadratic programming methods for parametric nonlinear optimization," *Comput. Optim. Appl.*, vol. 59, no. 3, pp. 475–509, Feb. 2014.

[38] G. Scutari and Y. Sun, *Parallel and distributed successive convex approximation methods for big-data optimization*. Springer-Verlag, 2018, pp. 141–308.

[39] J. Wen, K. Huang, S. Yang, and V. O. K. Li, "Cache-enabled heterogeneous cellular networks: Optimal tier-level content placement," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5939–5952, 2017.

[40] K. Li, C. Yang, Z. Chen, and M. Tao, "Optimization and analysis of probabilistic caching in N-tier heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1283–1297, 2018.

[41] D. Liu and C. Yang, "Caching policy toward maximal success probability and area spectral efficiency of cache-enabled HetNets," *IEEE Trans. Commun.*, vol. 65, no. 6, pp. 2699–2714, 2017.

[42] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *IEEE Int. Conf. Comput. Commun., (INFOCOM)*, March 1999, pp. 126–134.

[43] 3GPP, "Universal mobile telecommunications system (UMTS); radio frequency RF system scenarios," 3rd Generation Partnership Project (3GPP), Technical Report (TR), April 2017, version 14.0.0.

[44] S. N. Chiu, D. Stoyan, W. S. Kendall, and J. Meckee, *Stochastic Geometry and Its Applications*, 3rd ed. John Wiley & Sons, 2013.

[45] X. Xu and M. Tao, "Modeling, analysis, and optimization of caching in multi-antenna small-cell networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5454–5469, 2019.

APPENDIX

To prove Theorem 1, we need the following Lemmas.

Lemma 1. *Let $\{\boldsymbol{x}_i\}_i$ be the points of a homogeneous PPP Φ with intensity λ, and $S(\cdot)$ and $P(\cdot)$ be two real-valued functions on the state space of Φ. We have:*

$$\mathbb{E}\left\{\prod_{k\in\Phi} P(\boldsymbol{x}_k)\sum_{k\in\Phi} S(\boldsymbol{x}_k)\right\} = \lambda\iint_{\mathbb{R}^2} S(\boldsymbol{s})P(\boldsymbol{s})d\boldsymbol{s}\,\exp\left(\lambda\iint_{\mathbb{R}^2}(P(\boldsymbol{s})-1)d\boldsymbol{s}\right).$$

Proof. Let N denote the number of points of Φ, which follows a Poisson distribution. Suppose that the points are placed in the region A as a subspace of Cartesian space. Note that for a homogeneous PPP, the points can be considered to be independently and uniformly distributed over A [44]. We then have:

$$\begin{aligned}
\mathbb{E}\Big\{\prod_{k\in\Phi} P(\boldsymbol{x}_k)\sum_{k\in\Phi} S(\boldsymbol{x}_k)\Big\} &= \mathbb{E}_N\Big\{\mathbb{E}\Big\{\sum_{k=1}^{N} S(\boldsymbol{x}_k)P(\boldsymbol{x}_k)\prod_{i\neq k}^{N} P(\boldsymbol{x}_i)\Big|N\Big\}\Big\}\\
&\overset{(a)}{=} \mathbb{E}_N\left\{\sum_{n=1}^{N}\iint_A \frac{S(\boldsymbol{s})P(\boldsymbol{s})}{|A|}d\boldsymbol{s}\left(\iint_A\frac{P(\boldsymbol{s})}{|A|}d\boldsymbol{s}\right)^{N-1}\right\}\\
&= \iint_A \frac{S(\boldsymbol{s})P(\boldsymbol{s})}{|A|}d\boldsymbol{s}\;\mathbb{E}_N\left\{N\left(\iint_A\frac{P(\boldsymbol{s})}{|A|}d\boldsymbol{s}\right)^{N-1}\right\}\\
&\overset{(b)}{=} \iint_A \frac{S(\boldsymbol{s})P(\boldsymbol{s})}{|A|}d\boldsymbol{s}\,\lambda|A|\,\exp\big(-\lambda|A|\big)\,\exp\left(\lambda|A|\iint_A\frac{P(\boldsymbol{s})}{|A|}d\boldsymbol{s}\right)
\end{aligned}$$

where $|A|$ is the area of A, for (a) we consider that points are independently and uniformly distributed over A, and for (b) we use $N\sim\text{Pois}(\lambda|A|)$. By rearranging the last equation and letting $A=\mathbb{R}^2$, the statement follows. □

Lemma 2. *Assume $\{\boldsymbol{r}_i\}_{i\in\Phi}$ being the Poisson points located based on a PPP Φ, and consider Voronoi tessellation constructed by these points with $\mathcal{V}_0$ as the Voronoi cell associated with $\boldsymbol{r}_0$, then the quantity $\mathbb{1}(\boldsymbol{x}_i\in\mathcal{V}_0)$, for any location $\boldsymbol{x}\in\mathbb{R}^2$, can be expressed based on the following product of Φ:*

$$\mathbb{1}(\boldsymbol{x}\in\mathcal{V}_0) = \prod_{k\in\Phi\setminus\{0\}}\mathbb{1}\big(\boldsymbol{x}\in M_0(\boldsymbol{r}_k)\big),$$

where $\mathbb{1}(\cdot)$ is the indicator function and

$$M_0(\boldsymbol{r}_k) := \big\{\boldsymbol{r}'\in\mathbb{R}^2 \;\big|\; \|\boldsymbol{r}'-\boldsymbol{r}_0\|\le\|\boldsymbol{r}'-\boldsymbol{r}_k\|\big\},\qquad \textit{for } k\in\Phi\setminus\{0\}.$$

Proof. The statement follows based on the geometrical definition of the Voronoi tessellation of Φ and definition of $M_0(\cdot)$. □

We also need to compute the expectation of resources needed for a typical UE requesting from the SPUC component. Based on (7) and (8) and by defining $w_i=\dfrac{R}{\log(1+\gamma_i)}$, we get:

$$\begin{aligned}
\mathbb{E}_{\gamma_i}\left\{\frac{R}{\log(1+\gamma_i)}\mathbb{1}(\gamma_i\ge\gamma)\right\} &= \mathbb{E}_{w_i}\left\{\frac{R}{\log(1+\gamma_i)}\mathbb{1}(w_i\le w_t)\right\}\\
&= \int_0^{w_t} w\frac{d}{dw}\mathrm{F}_{w_i}(w)dw,
\end{aligned}$$

where $\mathrm{F}_{w_i}(w)$ is the CDF of r.v. w_i evaluated at w, Based on (7), we obtain:

$$\begin{aligned}
\mathrm{F}_{w_i}(w) &= 1-\mathbb{P}\left\{g_i\|\boldsymbol{x}_i\|^{-e}\le I_i\eta(w)\right\}\\
&\overset{(a)}{=} \sum_{l=1}^{L-n_z+1} f_{l,n_z}\mathbb{E}_{I_i}\left\{\exp(-l\xi\eta(w)I_i\|\boldsymbol{x}_i\|^e)\right\},
\end{aligned}$$

where (a) is obtained based on the gamma distribution of $g_i\sim\Gamma(L-n_z+1,1)$ and promising approximation used in [45], though for $L=1$ it is exact. We also have $f_{l,n_z}=(-1)^{l+1}\binom{L-n_z+1}{l}$ and $\eta(w)=2^{\frac{R}{w}}-1$. Considering that $I_i=\sum_{k\in\Phi_{\text{bs}}} g_k^i\|\boldsymbol{x}_i-\boldsymbol{r}_k\|^{-e}$ and $g_k^i\sim\Gamma(n_z,1)$ we can get:

$$\begin{aligned}
\mathbb{E}_{I_i}\left\{\exp\left(-l\xi\eta(w)I_i\|\boldsymbol{x}_i\|^e\right)\right\} &= \prod_{k\in\Phi_{\text{bs}}}\left(1+l\xi\eta(w)\frac{\|\boldsymbol{x}_i-\boldsymbol{r}_k\|^{-e}}{\|\boldsymbol{x}_i\|^{-e}}\right)^{-n_z}\\
&= \prod_{k\in\Phi_{\text{bs}}} P_{w,l}\left(\frac{\|\boldsymbol{x}_i-\boldsymbol{r}_k\|^{-e}}{\|\boldsymbol{x}_i\|^{-e}}\right),
\end{aligned}$$

where $P_{w,l}(\zeta):=\big(1+l\xi\eta(w)\zeta\big)^{-n_z}$. By using $\dfrac{d}{dw}\mathrm{F}_{w_i}(w)=\mathrm{F}_{w_i}(w)\dfrac{d}{dw}\log\left(\mathrm{F}_{w_i}(w)\right)$, and defining

$$S_{w,l}(\zeta) := -n_z\frac{d\eta(w)}{dw}\frac{l\xi\,\zeta}{1+l\xi\eta(w)\,\zeta},$$

we can obtain:

$$\begin{aligned}
\mathbb{E}_{\gamma_i}\left\{\frac{R}{\log(1+\gamma_i)}\mathbb{1}(\gamma_i\ge\gamma)\right\} &= \sum_{l=1}^{L-n_z+1} f_{l,n_z}\int_0^{w_t} w\\
&\times\sum_{k\in\Phi_{\text{bs}}} S_{w,l}\left(\frac{\|\boldsymbol{x}_i-\boldsymbol{r}_k\|^{-e}}{\|\boldsymbol{x}_i\|^{-e}}\right)\prod_{k\in\Phi_{\text{bs}}} P_{w,l}\left(\frac{\|\boldsymbol{x}_i-\boldsymbol{r}_k\|^{-e}}{\|\boldsymbol{x}_i\|^{-e}}\right)dw. \quad (28)
\end{aligned}$$

$$W' = \sum_{l=1}^{L-n_z+1} f_{l,n_z} \int_0^{w_t} w \mathbb{E}_{\Phi_{\rm ue}} \Big\{ \sum_{i\in\Phi_{\rm ue}} \lambda_{\rm bs} \iint_D Q_{w,l}\left(\frac{\|\boldsymbol{x}_i-\boldsymbol{r}\|^{-e}}{\|\boldsymbol{x}_i\|^{-e}}\right) \mathbb{1}\big(\boldsymbol{x}_i \in M_0(\boldsymbol{r})\big) r dr d\theta \\ \exp\left(\lambda_{\rm bs} \iint_D \Big(P_{w,l}\left(\frac{\|\boldsymbol{x}_i-\boldsymbol{r}\|^{-e}}{\|\boldsymbol{x}_i\|^{-e}}\right) \mathbb{1}\big(\boldsymbol{x}_i \in M_0(\boldsymbol{r})\big) - 1\Big) r dr d\theta\right)\Big\} dw, \tag{30}$$

Now, we can compute $W' = \mathbb{E}\left\{\sum_{i\in\mathcal{V}_0} w_i\right\}$ from (9). Without loss of generality, we assume that BS responsible in the cell $\mathcal{V}_0$ is located at origin, i.e., $\boldsymbol{r}_0 = \boldsymbol{0}$. Based on (7), we have:

$$\begin{aligned} W' &= \mathbb{E}_{\Phi_{\rm ue},\Phi_{\rm bs}} \mathbb{E}_{\boldsymbol{\gamma}} \left\{ \sum_{i\in\mathcal{V}_0} \frac{R}{\log(1+\gamma_i)} \mathbb{1}(\gamma_i \geq \gamma) \right\} \\ &= \mathbb{E}_{\Phi_{\rm ue},\Phi_{\rm bs}} \left\{ \sum_{i\in\Phi_{\rm ue}} \mathbb{E}_{\gamma_i} \left\{ \frac{R}{\log(1+\gamma_i)} \mathbb{1}(\gamma_i \geq \gamma) \right\} \mathbb{1}(\boldsymbol{x}_i \in \mathcal{V}_0) \right\} \\ &\overset{(a)}{=} \mathbb{E}_{\Phi_{\rm ue}} \Big\{ \sum_{i\in\Phi_{\rm ue}} \mathbb{E}_{\Phi_{\rm bs}} \Big\{ \sum_l f_{l,n_z} \int_0^{w_t} w \sum_{k\in\Phi_{\rm bs}\setminus\{0\}} S_{w,l}\left(\frac{\|\boldsymbol{x}_i-\boldsymbol{r}_k\|^{-e}}{\|\boldsymbol{x}_i\|^{-e}}\right) \\ &\quad \times \prod_{k\in\Phi_{\rm bs}\setminus\{0\}} P_{w,l}\left(\frac{\|\boldsymbol{x}_i-\boldsymbol{r}_k\|^{-e}}{\|\boldsymbol{x}_i\|^{-e}}\right) \mathbb{1}\big(\boldsymbol{x}_i \in M_0(\boldsymbol{r}_k)\big) dw \Big\} \Big\} \end{aligned} \tag{29}$$

where for (a) we use (28) and Lemma 2. In (29) we are dealing with a reduced-palm process as $k \in \Phi_{\rm bs}\setminus\{0\}$. Considering that the distribution of reduced-palm process is equal to the distribution of original process [31], and based on Lemma 1, we obtain (30) being written in the top of this page. Where $D = \{(r,\theta)|\ 0 \leq r \leq \infty, 0 \leq \theta \leq 2\pi\}$ and $Q_{w,l}(\cdot) := S_{w,l}(\cdot)P_{w,l}(\cdot)$. By considering:

$$\mathbb{1}\big(\boldsymbol{x}_i \in M_0(\boldsymbol{r})\big) = B_c\big(\boldsymbol{x}_i, \|\boldsymbol{x}_i\|\big),$$

where $B_c(\boldsymbol{x}, \|\boldsymbol{x}\|)$ denotes the region outside the disk of radius $\|\boldsymbol{x}\|$ centered at $\boldsymbol{x}$, we then get:

$$W' \overset{}{=} \sum_{l=1}^{L-n_z+1} f_{l,n_z} \int_0^{w_t} w \lambda_{\rm bs}\lambda_{\rm ue} \iint_D \left\{ \iint_{B_c(\boldsymbol{x},\|\boldsymbol{x}\|)} Q_{w,l}\left(\frac{\|\boldsymbol{x}-\boldsymbol{r}\|^{-e}}{\|\boldsymbol{x}\|^{-e}}\right) r dr d\theta \right. \\ \left. \exp\left(\lambda_{\rm bs} \iint_{B_c(\boldsymbol{x},\|\boldsymbol{x}\|)} P_{w,l}\left(\frac{\|\boldsymbol{x}-\boldsymbol{r}\|^{-e}}{\|\boldsymbol{x}\|^{-e}}\right) r dr d\theta - \lambda_{\rm bs} \iint_D r dr d\theta\right)\right\} x dx d\theta' dw, \tag{31}$$

where $x = \|\boldsymbol{x}\|$. Note that (31) obtained by considering the probability generating functional (PGFL) property of PPP $\Phi_{\rm ue}$ [31]. Now, we consider change of variables $\boldsymbol{\rho} = \boldsymbol{x} - \boldsymbol{r}$, and then $\boldsymbol{q} = z\boldsymbol{\rho}$ to obtain:

$$\begin{aligned} W' &= \sum_{l=1}^{L-n_z+1} f_{l,n_z} \int_0^{w_t} w \lambda_{\rm bs}\lambda_{\rm ue} \iint_D \Big\{ 2\pi \int_1^\infty Q_{w,l}\left(z^{-e}\right) q^2 z dz \\ &\quad \exp\Big(2\pi\lambda_{\rm bs} q^2 \Big(\int_1^\infty P_{w,l}\left(z^{-e}\right) z dz - \int_0^\infty z dz \Big)\Big)\Big\} q dq dw \\ &\overset{(a)}{=} \sum_{l=1}^{L-n_z+1} f_{l,n_z} \int_0^{w_t} \Big(\frac{w\lambda_{\rm ue}}{2\lambda_{\rm bs}} \int_1^\infty Q_{w,l}(z^{-e}) z dz\, \Omega_l^2(w)\Big) dw \\ &= \frac{\lambda_{\rm ue}}{2\lambda_{\rm bs}} \sum_{l=1}^{L-n_z+1} f_{l,n_z} \int_0^{w_t} w \frac{d}{dw}\, \Omega_l(w) dw, \end{aligned}$$

where

$$\Omega_l(w) = \frac{1}{\frac{1}{2} - \int_1^\infty \left(P_{w,l}(z^{-e}) - 1\right) z dz},$$

$\rho = \|\boldsymbol{\rho}\|$, $q = \|\boldsymbol{q}\|$, $\frac{d}{dw} P_{w,l}(\cdot) = Q_{w,l}(\cdot)$ and for (a) we use $\int_0^\infty q^3 \exp(\phi_0 q^2) dq = \frac{1}{2\phi_0^2}$ for $\phi_0 < 0$. By computing the inner integral, the statement follows.

Now, we compute $\mathcal{O}^{\rm UC}(\gamma)$, i.e., the outage probability of a typical UE served by the SPUC component. Without loss of generality, we assume that the UE is located at origin. According to the truncated SINR policy (8), this UE is in outage if $\mathbb{P}(\gamma_k^{\rm UC} \leq \gamma)$. Hence, we get:

$$\begin{aligned} \mathcal{O}^{\rm UC}(\gamma) &= \mathbb{P}(\gamma_k^{\rm UC} \leq \gamma) = \mathbb{P}(g_0 \|\boldsymbol{r}_0\|^{-e} \leq \gamma I_0) \\ &\overset{(a)}{=} 1 - \sum_{l=1}^{L+n_z-1} f_{l,n_z}\, \mathbb{E}_{I_0,\Phi_{\rm bs}\setminus\{0\},\boldsymbol{r}_0} \left\{ \exp(-l\xi\gamma I_0 \|\boldsymbol{r}_0\|^e) \right\}, \end{aligned}$$

where $\boldsymbol{r}_0$ is the location of nearest BS w.r.t. the UE, $I_0 = \sum_{j\in\Phi_{\rm bs}\setminus\{0\}} g_j \|\boldsymbol{r}_j\|^{-e}$ is the interfering term and (a) is obtained considering $g_0 \sim \Gamma(L-n_z+1, 1)$. We then get:

$$\begin{aligned} \mathcal{O}^{\rm UC}(\gamma) &= 1 - \sum_{l=1}^{L+n_z-1} f_{l,n_z} \mathbb{E}\Big\{ \prod_{j\in\Phi_{\rm bs}\setminus\{0\}} \exp\Big(-l\xi\gamma g_j \frac{\|\boldsymbol{r}_j\|^{-e}}{\|\boldsymbol{r}_0\|^{-e}}\Big)\Big\} \\ &\overset{(a)}{=} 1 - \sum_{l=1}^{L+n_z-1} f_{l,n_z} \mathbb{E}\Big\{ \prod_{j\in\Phi_{\rm bs}\setminus\{0\}} \Big(1 - l\xi\gamma \frac{\|\boldsymbol{r}_j\|^{-e}}{\|\boldsymbol{r}_0\|^{-e}}\Big)^{-n_z}\Big\} \\ &\overset{(b)}{=} 1 - \sum_{l=1}^{L+n_z-1} f_{l,n_z} \mathbb{E}\Big\{ \exp\Big(2\pi\lambda_{\rm bs} \int_{r_0}^\infty \Big((1 - l\,\xi\,\gamma\, r_0^e\, r^{-e})^{-n_z} - 1\Big) r dr\Big)\Big\} \\ &= 1 - \sum_{l=1}^{L+n_z-1} f_{l,n_z} \mathbb{E}\left\{ \exp\left(\pi\lambda_{\rm bs} r_0^2 \Big(1 - \Theta_{l,n_z}\Big)\right)\right\} \\ &\overset{(c)}{=} 1 - \sum_{l=1}^{L+n_z-1} f_{l,n_z} \int_0^\infty 2\pi\lambda_{\rm bs} r_0 \exp\big(\pi\lambda_{\rm bs} r_0^2 (1-\Theta_{l,n_z})\big) \exp\big(-\pi\lambda_{\rm bs} r_0^2\big) dr_0 \\ &= 1 - \sum_{l=1}^{L-n_z+1} \frac{f_{l,n_z}}{\Theta_{l,n_z}}, \end{aligned} \tag{32}$$

where $\Theta_{l,n_z} = {}_2F_1\Big(-\frac{2}{e}, n_z, 1-\frac{2}{e}, -l\,\xi\,\gamma\Big)$, $r_0 = \|\boldsymbol{r}_0\|$, (a) is obtained by expectation w.r.t. $g_j \sim \Gamma(n_z, 1)$, we use PGFL property of PPP $\Phi_{\rm bs\setminus\{0\}}$ for (b), and for (c) we regard the PDF of r_0: $f(r_0) = 2\pi\lambda_{\rm bs} r_0 \exp(-\pi\lambda_{\rm bs} r_0^2)$.

Mohsen Amidzadeh received the B.Sc. and M.Sc. degrees in electrical engineering from Sharif University of Technology, Tehran, Iran. He is currently pursuing a PhD degree at department of information and communications, Aalto University, Finland. He completed a sabbatical program with the University of Alberta, Edmonton, Canada, in 2018. His current research interests include next-generation cellular networks, wireless communications, machine learning, optimization problems, and estimation theory.

Giuseppe Caire (S '92 – M '94 – SM '03 – F '05) was born in Torino in 1965. He received a B.Sc. in Electrical Engineering from Politecnico di Torino in 1990, an M.Sc. in Electrical Engineering from Princeton University in 1992, and a Ph.D. from Politecnico di Torino in 1994. He has been a post-doctoral research fellow with the European Space Agency (ESTEC, Noordwijk, The Netherlands) in 1994-1995, Assistant Professor in Telecommunications at the Politecnico di Torino, Associate Professor at the University of Parma, Italy, Professor with the Department of Mobile Communications at the Eurecom Institute, Sophia-Antipolis, France, a Professor of Electrical Engineering with the Viterbi School of Engineering, University of Southern California, Los Angeles, and he is currently an Alexander von Humboldt Professor with the Faculty of Electrical Engineering and Computer Science at the Technical University of Berlin, Germany.

He received the Jack Neubauer Best System Paper Award from the IEEE Vehicular Technology Society in 2003, the IEEE Communications Society and Information Theory Society Joint Paper Award in 2004 and in 2011, the Okawa Research Award in 2006, the Alexander von Humboldt Professorship in 2014, the Vodafone Innovation Prize in 2015, an ERC Advanced Grant in 2018, the Leonard G. Abraham Prize for best IEEE JSAC paper in 2019, the IEEE Communications Society Edwin Howard Armstrong Achievement Award in 2020, the 2021 Leibniz Prize of the German National Science Foundation (DFG), and the CTTC Technical Achievement Award of the IEEE Communications Society in 2023. Giuseppe Caire is a Fellow of IEEE since 2005. He has served in the Board of Governors of the IEEE Information Theory Society from 2004 to 2007, and as officer from 2008 to 2013. He was President of the IEEE Information Theory Society in 2011. His main research interests are in the field of communications theory, information theory, channel and source coding with particular focus on wireless communications.

Olav Tirkkonen (Fellow'23) is full professor in communication theory at Aalto University, Finland, where he has held a faculty position since 2006. He received his M.Sc. and Ph.D. degrees in theoretical physics from Helsinki University of Technology in 1990 and 1994, respectively. After post-doctoral positions at UBC, Vancouver, Canada, and NORDITA, Copenhagen, Denmark, he was with Nokia Research Center (NRC), Helsinki, Finland from 1999 to 2010. In 2016-2017 he was Visiting Associate Professor at Cornell University, Ithaca, NY, USA. He has published some 300 papers, and is the inventor of some 85 families of patents and patent applications which include 1% of all patents declared essential for the first standardized version of 4G LTE. His current research interests are in coding for random access and quantization, quantum computation, and machine learning for cellular networks. He served as General Chair of 2022 IEEE International Symposium on Information Theory, and is a member of the Executive Editorial Committee of IEEE Transactions on Wireless Communications.