Amidzade, Mohsen; Al-Tous, Hanan; Tirkkonen, Olav; Caire, Giuseppe

Caching in Cellular Networks Based on Multipoint Multicast Transmissions

Published in:
IEEE Transactions on Wireless Communications

DOI:
10.1109/TWC.2022.3211416

Published: 01/04/2023

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

Please cite the original version:
Amidzade, M., Al-Tous, H., Tirkkonen, O., & Caire, G. (2023). Caching in Cellular Networks Based on Multipoint Multicast Transmissions. *IEEE Transactions on Wireless Communications*, *22*(4), 2393-2408. https://doi.org/10.1109/TWC.2022.3211416

Caching in Cellular Networks Based on Multipoint Multicast Transmissions

Mohsen Amidzadeh*, Hanan Al-Tous*, Giuseppe Caire† and Olav Tirkkonen*
*Department of Communications and Networking, Aalto University, Finland
†Communications and Information Theory Chair, TU Berlin, Germany
{mohsen.amidzade, hanan.al-tous, olav.tirkkonen}@aalto.fi, caire@tu-berlin.de

***Abstract*—We consider cellular network caching with network-wide Orthogonal Multipoint Multicast (OMPMC) delivery. We apply a probabilistic model for content placement at the Base Stations (BSs). Content is delivered with multipoint multicast operating in file-specific orthogonal resources: all BSs caching a distinct file synchronously multicast it to requesting users in a dedicated resource. For a network modeled as a Poisson Point Process (PPP), an expression for the outage probability is derived. The outage-minimizing cache policy is found from a joint constrained optimization problem over cache placement and resource allocation. We devise principles by which the solution in one propagation environment can be generalized to another. To reduce computational complexity, we obtain a sub-optimal solution based on convex relaxation. We obtain an upper bound of the gap between the optimal and sub-optimal solutions. We compare the outage performance of OMPMC with delivery polices from the literature. Simulation results show that exploiting OMPMC with optimal cache placement and resource allocation outperforms single point cache delivery policies with a wide margin.**

***Index Terms*—Wireless caching, cache placement and delivery policy, multipoint multicast transmission, resource allocation, convex relaxation, gap analysis.**

I. Introduction

Edge caching can be applied to reduce backhaul link congestion and to reduce latency in cellular networks [1], [2]. In wireless caching, two successive phases are considered; cache placement and cache delivery [3]. In the placement phase, content is proactively stored at the network edge during off-peak hours, based on estimated user preferences. In the delivery phase, cached content is transmitted towards requesting users. In an optimal caching policy, cache placement and delivery should be jointly optimized.

Cache placement can be based on either a probabilistic [4]–[16] or deterministic [17]–[21] method. A probabilistic method is scalable and can be applied to the large networks. In [4], a probabilistic approach was used to place the contents at the Base-Stations (BSs), with BSs independently storing content at their caches according to the same probability distribution. In [5], [9], probabilistic placement is utilized for Heterogeneous Networks (HetNets) to design a cache policy. Successful delivery probability is optimized for a HetNet in [11] to optimally design a probabilistic tier-level cache placement. In [12], a probabilistic cache placement is optimized for a cache-enabled HetNet based on a tier-specific bandwidth allocation.

In the cache delivery phase, it is important to distinguish between unicast and multicast approaches, as well as between approaches where a single point (the closest caching BS) delivers a file, and multipoint approaches, where a user downloading a file simultaneously receives transmissions from multiple BSs.

Single-point unicast (SPUC) delivery has been considered in [4]–[12], [19] for on-demand cache delivery. In [6], the requesting UEs are satisfied by the nearest BS and a dynamic network architecture is considered, where BSs are only active on demand by UE. In [9], a SPUC scheme is applied where a HetNet, with ZF-beamforming BSs and cache-enabled helper-nodes, is considered. In [10], a HetNet applying SPUC is considered. The requesting UE is served by a caching BS that has the maximum received signal power. SPUC is utilized for an Unmanned Aerial Vehicle (UAV)-aided network with the caching capability to alleviate the backhaul load based on an

This work was funded in part by the Academy of Finland (grant 319058). The work of Giuseppe Caire was partially funded by the European Research Council under the ERC Advanced Grant N. 789190, CARENET.

interference alignment [19]. To secure the transmission, an eavesdropper disrupting scheme is exploited.

Single-point multicast (SPMC) approaches are considered in [13]–[15], [17], [18]. In these, each caching BS multicasts or broadcasts different files towards requesting UEs using a multiple access technique. In [13]–[15], the authors consider probabilistic caching in a HetNet where each BS multicasts k files in the pre-assigned resources each spanning $1/k$ of total bandwidth. In [17], the authors consider a multicast sparse beamforming based on a deterministic cache placement strategy. In [18], a broadcast transmission is applied for an UAV-assisted network with caching capability. The trajectory of the UAV relay is optimized to secure the transmission against the eavesdroppers.

Deterministic caching multipoint unicast (MPUC) schemes are considered in [20], [21] for on-demand cache delivery. In [20], the caching BSs transmit a cached file towards the requesting UEs over a dedicated resource. Resources in the network are orthogonalized between UEs. The BSs have a single antenna, and CSI is not used for collaborative beamforming. In [21], Coordinated Multipoint (CoMP) transmission with Zero-Forcing beamforming is used. The CSI between UEs and a set of UE-serving BSs is assumed known. For each UE, unicast CoMP transmissions from the serving set of BSs are applied which yields unavoidable multiple access interference.

Wireless content delivery based on multipoint broadcast transmissions are applied in Digital Terrestrial TV Broadcasting systems [22], while multipoint multicast (MPMC) content delivery has been applied in the Long Term Evolution (LTE) system in the context of enhanced-MBMS [23], [24]. In a multi-cell transmission mode, all serving BSs broadcast the same file into the air in an Single-Frequency-Network (SFN) configuration. As such, the file is concurrently transmitted over the same frequency bandwidth within the whole network.

An non-SFN MPMC scheme has been considered together with coded caching at the user end in [16]. In this approach, each transmitter transmits different Minimum Distance Separable (MDS) coded content fragments, and each UE separately decodes transmissions from multiple edge nodes. In this scheme, the only collaboration between BSs comes from the MDS coding, and the BSs are not equipped with caches.

Stochastic geometry [25]–[27], is a versatile tool for modeling and analyzing wireless network, both for homogeneous and HetNets. In the context of caching at the wireless edge, PPPs have been used to model deployment of BSs and locations of User Equipment (UE) in [7]–[9], [15]. In [9], two independent PPPs are exploited to model the deployment of each tier of the HetNets. Aiming to find an optimal cache placement policy, independent PPPs are exploited in [15] to obtain an expression for the successful transmission probability in a HetNet. In [7], the coverage probability is approximated using PPP for the multi-antenna small-cell networks to design an optimal cache placement policy. In [8], the authors leverage PPP to optimize the minimum of the cache hit rates of different request-related categorized UEs.

In this paper[1], we propose to use SFN-configured MPMC for edge caching cellular networks. In the large networks with multiple UEs requesting the same file, single point delivery suffers from co-channel interference. This interference can be removed by SFN-based multicast transmissions which jointly address all request for a file in the network. Therefore, in contrast to SPUC and SPMC schemes with co-channel interference between cells, we consider Orthogonal MPMC (OMPMC) delivery based on a Network-Wide (NWW) *file-specific resource allocation* and cache placement strategy. For each file, the network constructs an SFN operating in orthogonal frequency resources. In contrast to the MPUC scheme of [20], we orthogonalize resources between files, not users. Our approach is thus complementary to [20]; the deterministic cache placement approach of [20] applies in small networks where the population of users is small as compared to the file library size, while our approach applies in large networks with a large population of users as compared to the library size. To get an analytic handle on the problem, we model the large network using PPP for both users and BSs. In contrast to [16], we consider SFN configuration and caching at the edge of the network, not at the UEs, and assume UEs cannot separate signals from different BSs.

The main contributions of this paper are as follows.

- *Notion of OMPMC caching in SFN configuration*: We consider a cache-enabled network applying an OMPMC delivery scheme with NWW resource allocation, creating multiple file-specific SFNs. We then analyze the overall outage probability of such a scheme, formulate the cache policy as a joint constrained optimization problem over cache placement and resource allocation, and characterize the optimal solution of the formulated problem.

[1]Early results of this work were presented in [28]

- *Finding optimal cache solutions and their generalizations*: We derive generalizing properties that enable us to find the optimal solution for one set of parameters, if we know the solution for another set. Specifically, we generalize solutions to different BS capacities, different file popularities, and different path-loss environments.
- *Devising efficient sub-optimal caching solutions*: By relaxing the optimization problem, sub-optimal solutions are found. We evaluate their computational complexity and analyze the relaxation gap.
- *Numerical evaluation of OMPMC outage performance*: We compare OMPMC to cache policies from the literature and heuristic solutions, using numerical simulations in different settings.

The remainder of this paper is organized as follows. In Section II, the system model is introduced. In Section III, the cache policy problem under OMPMC scheme is formulated. We analyze the optimal cache policy and devise an algorithm to obtain the optimal policy in Section IV. Cache policies from literature are addressed for comparison in Section V. The simulation results are presented and discussed in Section VI. Finally, Section VII concludes the paper.

Notations: In this paper, we use bold-face lower-case and bold-face uppercase letters to indicate vectors and matrices, respectively. Further, $\boldsymbol{A}^\top$ is matrix transpose, and we indicate the inverse of the function $f(x)$ by $f^{-1}(x)$ and its derivative with respect to x by $D_x f(x)$ or $f'(x)$. Moreover, $\nabla g(\boldsymbol{a})$ represents the Gradient column vector of the function $g(\cdot)$ with respect to $\boldsymbol{a}$, $D_{x,y}h(x,y)$ indicates the derivative of the function $h(x,y)$ with respect to x and y. We show the Euclidean L^2 norm by $\|\cdot\|$ and components of a n-dimensional column vector $\boldsymbol{b}$ using the notation $\boldsymbol{b} = [b_1, \dots, b_n]^\top$. The m-th element of the vector $\boldsymbol{b}$ is denoted as $[\boldsymbol{b}]_m = b_m$, and $\{b_m\}_1^p$ collects the components of vector $\boldsymbol{b}$ from $m = 1$ to $m = p$.

II. System Model

We consider a cellular network populated with cache-equipped BSs and UEs. A content library containing N files is considered from which the BSs fetch files to store at their caches. We assume that the popularity of files is known. The popularity stands for the probability that a distinct file is requested by UEs. We consider the Zipf distribution for the file popularity [29]. Accordingly, f_n denotes the probability that file n is requested, and $f_n = n^{-\theta}/\sum_{j=1}^{N} j^{-\theta}$ for $n = 1, \dots, N$, where θ denotes the skewness of the Zipf distribution. Without loss of generality, we assume that all files have the same size, normalized to 1 bit. The content placement at BSs cache is performed using a probabilistic approach, for this, each BS independently caches the files based on common caching weights. For the cache delivery, the network employs OMPMC transmission scheme. Based on this strategy, all BSs that cache a specific file, broadcast it towards requesting UEs. It is very important that the transmitting BSs are well synchronized, within a fraction of a cyclic prefix. With current technology, it is possible to achieve such high precision in network synchronization [30]. Accordingly, we shall neglect synchronization errors. To remove the interference during transmission of different files, a network-wide orthogonal transmission is applied. For this, a request for file n is served by all BSs caching this file in a fraction resource w_n. Two independent PPPs are used to model the deployment of BSs and the location of UEs. We denote the PPP modeling BSs deployment with Φ and its intensity with λ.

A. Cache Placement Policy

The BSs are equipped with caches of limited storage capacity such that they can store at most L files. In addition, BSs independently and randomly cache the files according to a common probabilistic approach as in [4], [5]. More specifically, to store file n, BSs refer to the caching weight $\{q_n\}_1^N$, $q_n \leq 1$, where q_n indicates the probability that file n is stored at a randomly selected BS. To respect the capacity of BSs, we have $\sum_{n=1}^{N} q_n = L$. According to the thinning property of Φ, files are cached at BSs based on another PPP with intensity $\{\lambda q_n\}_1^N$ [26].

B. Cache Delivery Policy

The network operates according to a time-slotted fashion. For each time-slot, a spatial distribution of UEs exists that request contents based on the files popularity. The network applies the OMPMC transmission scheme to serve the requesting UEs at each time-slot.

For a distinct file request, all BSs caching the file are exploited to simultaneously broadcast it in a dedicated resource. Moreover, an orthogonal NWW allocation of resources to files is leveraged. For this, no transmissions are occurred in the radio resource reserved for a distinct file throughout the whole network except the multicast transmission of this file. The fraction of resource dedicated for file n is denoted by w_n, where $\sum_{n=1}^{N} w_n = 1$.

We assume that the average transmission power of all BSs is the same and denoted by $p_{\rm tx}$. Moreover, each BS allocates a fractional power $p_{\rm tx} w_n$ to broadcast file n in dedicated resource w_n. In this case, the transmitting Signal-to-Noise-Ratio (SNR) related to file n is computed as $\gamma_n^{(\rm tx)} = \frac{w_n p_{\rm tx}}{w_n W N_0}$, where N_0 and W are the noise spectral density and total transmission bandwidth, respectively. Note that $\gamma_n^{(\rm tx)} = \frac{p_{\rm tx}}{W N_0}$ is independent of file index n, so we denote it by $\gamma_{\rm tx}$.

For file delivery, the wide-band macro-diversity transmission/reception is assumed. Each UE demanding a distinct file receives the corresponding signal over a multipath channel, which is the aggregation of transmitted signals from all BSs caching the file. For this, the Orthogonal Frequency Division Multiplexing (OFDM) SFN transmissions, leveraged in Digital Video Broadcasting [22], with sufficiently long Cyclic Prefix (CP), can be considered.

C. System Performance

Any UE requesting a file will be fulfilled by the network through OMPMC scheme. As declared, the file delivery is considered to be performed by the use of a NWW orthogonal and OFDM-based transmission, and due to this orthogonality, there will not be any co-channel interference. Further, it is assumed that each receiver is able to completely equalize the data without being affected by Inter-Carrier-Interference (ICI) or Inter-Block-Interfrence (IBI). The receiver performance in the presence of multiple transmitter and in an OFDM-based transmission is analyzed in [31]. The SNR of UE k demanding file n through OMPMC scheme is:

$$\gamma_{k,n} = \gamma_{\rm ref} \sum_{j \in \Phi_n} |h_{j,k}|^2 \|\boldsymbol{x}_j - \boldsymbol{r}_k\|^{-\beta}, \tag{1}$$

where Φ_n stands for the set of BSs storing file n, $h_{j,k}$ is the channel coefficient between UE k and BS j, $\boldsymbol{x}_j$ and $\boldsymbol{r}_k$ are the locations of BS j and UE k, respectively, $\beta > 2$ is the path-loss exponent related to the standard distance-dependent path-loss model, and $\gamma_{\rm ref} = w_n p_{\rm ref} / w_n W N_0 = p_{\rm ref}/W N_0$ is a reference SNR. Here $p_{\rm ref} = p_{\rm tx} x_{\rm ref}^{-\beta}$ is the average received power measured at a reference distance $x_{\rm ref}$ from a BS. Hence, we get $\gamma_{\rm ref} = \gamma_{\rm tx} x_{\rm ref}^{-\beta}$. Hereafter, we use the notations $p_{\rm ref}(\beta)$ and $\gamma_{\rm ref}(\beta)$ to emphasize the dependency of them on β.

The maximum achievable rate for an error-free transmission is based on the channel capacity at the presence of Additive-White-Gaussian-Noise (AWGN) channel. The allocated resources and the SNR associated with this transmission determine this rate. There exists a rate threshold $R_{\rm th}$ for a reliable transmission such that a UE experiences an outage if the maximum rate is less than this threshold. Since all files are assumed of the same size, we have: $\alpha = R_{\rm th}/W$. Hence, for library with size N, the resources available for transmitting each bit in the library is proportional to $1/(\alpha N)$. Accordingly, the outage probability for UE k requesting file n is

$$\mathcal{O}_{n,k} = \mathsf{Pr}\big(w_n \log_2(1 + \gamma_{k,n}) \le \alpha\big). \tag{2}$$

The outage probability $\mathcal{O}_{n,k}$ can be formulated using the channel gain threshold $\eta_n(\beta)$, as follows

$$\mathcal{O}_{n,k} = \mathsf{Pr}\left(\frac{\gamma_{k,n}}{\gamma_{\rm ref}(\beta)} \le \eta_n(\beta)\right), \tag{3}$$

where

$$\eta_n(\beta) = (2^{\alpha/w_n} - 1)/\gamma_{\rm ref}(\beta). \tag{4}$$

Note that the notation $\eta_n(\beta)$ is used to emphasize its dependence on β due to the variable $\gamma_{\rm ref}(\beta)$.

III. Overall Outage Probability for OMPMC

Having specified the outage probability associated with a file and a UE, we intend to define the overall outage probability of all file requests. We leverage it as the system metric to assess the performance of the cache policy. According to Slivnyak-Mecke theorem [26], for stationary and homogeneous PPPs, it suffices to compute the SNR for a UE located at the origin. For this UE, without loss of generality we can set $\gamma_n = \gamma_{0,n}$, $h_j = h_{j,0}$, $\boldsymbol{r}_0 = \mathbf{0}$ and $\mathcal{O}_n = \mathcal{O}_{n,0}$. Considering that files are requested based on the file popularity $\{f_n\}_1^N$, the overall outage probability is:

$$\mathcal{O}_{\rm tot} = \sum_{n=1}^{N} f_n \mathcal{O}_n, \tag{5}$$

where

$$\mathcal{O}_n = \mathsf{Pr}\Big(\underbrace{\sum_{j \in \Phi_n} |h_j|^2 \|\boldsymbol{x}_j\|^{-\beta}}_{z} \le \eta_n(\beta)\Big). \tag{6}$$

Thus, the outage probability of file n can be perceived as the Cumulative Distribution Function (CDF) of the random variable z at a value $\eta_n(\beta)$.

Proposition 1. *Assume OMPMC in a network with BSs distributed according to a homogenous PPP with intensity λ, with a propagation environment characterized by path-loss exponent β, and file $n = 1, \ldots, N$ with popularity f_n having* *caching weight q_n and bandwidth allocation w_n. The overall outage probability is then*

$$\mathcal{O}_{\rm tot} = \frac{2}{\pi}\sum_{n=1}^{N} f_n \int_0^\infty \left\{ \frac{1}{w}\cos\left(\frac{\pi^2\lambda q_n}{\beta\cos(\pi/\beta)}\left(\frac{w}{\eta_n(\beta)}\right)^{2/\beta}\right) \exp\left(-\frac{\pi^2\lambda q_n}{\beta\sin(\pi/\beta)}\left(\frac{w}{\eta_n(\beta)}\right)^{2/\beta}\right)\sin(w)\right\} dw.$$

Proof: Based on [25], we can compute the characteristic function $\phi_z(w) = \mathbb{E}\{e^{iwz}\}$ of the random variable z as

$$\phi_z(w) = \exp\left(\pi^2\lambda q_n \frac{w^{2/\beta}}{\beta}\left(\frac{i}{\cos(\pi/\beta)} - \frac{1}{\sin(\pi/\beta)}\right)\right).$$

Using Gil-Pelaez theorem, considering $\mathcal{O}_n = 0$ for $\eta_n(\beta) < 0$, and (5), the statement follows. ■

Note that in Proposition 1, the overall outage probability depends on a specific combination of caching weight and channel gain threshold, i.e., $q_n\eta_n(\beta)^{-2/\beta}$, and not on both separately.

IV. Optimal Cache Policy for OMPMC

A. Problem Formation

We exploit the overall outage probability as a performance metric to jointly optimize the cache placement and cache delivery through OMPMC scheme. By defining the set of library files as $S_N = \{1, \ldots, N\}$, the optimal cache policy for OMPMC scheme can be expressed as:

$$P_{(\beta)}: \quad \begin{aligned} &\min_{\{q_n\}_1^N, \{w_n\}_1^N} \sum_{n=1}^{N} f_n \mathcal{O}(w_n, q_n), \\ &\text{s.t.} \begin{cases} \sum_{n=1}^{N} w_n \le 1, & \sum_{n=1}^{N} q_n \le L, \\ 0 \le w_n \le 1, & 0 \le q_n \le 1, \quad \text{for } n \in S_N, \end{cases} \end{aligned} \tag{7}$$

where

$$\mathcal{O}(w_n, q_n) = \mathcal{O}_n = \frac{2}{\pi}\int_0^\infty \frac{1}{w}\cos\left(\frac{\lambda_\beta q_n}{\cos(\pi/\beta)}\left(\frac{w}{\eta_n(\beta)}\right)^{2/\beta}\right) \exp\left(-\frac{\lambda_\beta q_n}{\sin(\pi/\beta)}\left(\frac{w}{\eta_n(\beta)}\right)^{2/\beta}\right)\sin(w)\, dw,$$

and $\lambda_\beta = \dfrac{\pi^2\lambda}{\beta}$. We used the notation $P_{(\beta)}$ to emphasize the dependence of this problem on the path-loss parameter β.

Proposition 2. *Problem $P_{(\beta)}$, for all n for which $q_n > 0$, $w_n > 0$, is convex if:*

- $D_{q_nq_n}\mathcal{O}_n\ D_{w_nw_n}\mathcal{O}_n - D^2_{w_nq_n}\mathcal{O}_n \ge 0$ *for $\beta \ge 4$, or*
- $D_{q_nq_n}\mathcal{O}_n\ D_{w_nw_n}\mathcal{O}_n - D^2_{w_nq_n}\mathcal{O}_n \ge 0$ *and* $D_{q_nq_n}\mathcal{O}_n \ge 0$ *for $2 < \beta < 4$.*

Proof: It can be shown using the *leading principal minor* of the Hessian matrix [32]. Note that the constraints of $P_{(\beta)}$ are convex, so $P_{(\beta)}$ is a convex problem if the objective is a convex function. Considering the objective as a positive weighted sum of the functions $\mathcal{O}(w_n, q_n)$ for $n \in S_N$, it suffices to discuss the convexity of $\mathcal{O}(w_n, q_n)$. Let $\boldsymbol{H}$ denote the Hessian matrix of this function with respect to (w.r.t.) w_n and q_n. Its leading principal minor of order k, denoted by D_k, for $k = 1, 2$, can then be computed. Consequently, we have:

$$D_1 = D_{q_nq_n}\mathcal{O}_n \text{ and } D_2 = D_{q_nq_n}\mathcal{O}_n\, D_{w_nw_n}\mathcal{O}_n - D^2_{w_nq_n}\mathcal{O}_n.$$

Therefore, $P_{(\beta)}$ is convex, if $D_1 \ge 0$ and $D_2 \ge 0$. However, for $\beta \ge 4$, it has been shown, in the Appendix A, that D_1 is always positive. Hence, for $\beta \ge 4$ the convexity depends only on the positiveness of D_2. ■

Hereafter, we denote the optimal solution of (7) by $\{w_n^*\}_1^N$ and $\{q_n^*\}_1^N$. Some crucial properties of $P_{(\beta)}$ are:

- **Property A.1**: If $q_n = 0$, then outage probability for file n will be one regardless of the value of w_n, i.e., $\mathcal{O}(w_n, 0) = 1$. Likewise, if $w_n = 0$ then $\mathcal{O}(0, q_n) = 1$ regardless of value of q_n.
- **Property A.2**: $\mathcal{O}(w_n, q_n)$ is a monotonically decreasing function w.r.t. w_n and q_n.

  According to (6), the outage probability of file n, being cached with probability q_n and associated with channel threshold $\eta_n(\beta)$, is $F_{\gamma_n}\big(\gamma_{\rm ref}(\beta)\eta_n(\beta)\big)$, where γ_n is the SNR of file n being cached with probability q_n and $F_{\gamma_n}(\cdot)$ is the CDF of γ_n. On the other hand, based on Proposition 1, $\mathcal{O}_n$ merely depends on $q_n\eta_n(\beta)^{-2/\beta}$, not on q_n and $\eta_n(\beta)$ separately. Hence, we can re-express it as $\mathcal{O}_n = F_{\gamma'_n}\big(\gamma_{\rm ref}(\beta)\eta_n(\beta)q_n^{-\beta/2}\big)$, which is the outage probability for a case with channel threshold $\eta_n(\beta)q_n^{-\beta/2}$, SNR being γ'_n and file n being cached with probability 1. Now, we can compute:

  $$-D_{q_n}\mathcal{O}_n = \frac{\beta\,\gamma_{\rm ref}(\beta)\,\eta_n(\beta)}{2q_n^{\frac{\beta}{2}+1}}\,\mathfrak{f}_{\gamma'_n}\left(\frac{\gamma_{\rm ref}(\beta)\eta_n(\beta)}{q_n^{\frac{\beta}{2}}}\right) > 0,$$

  where $\mathfrak{f}_{\gamma'_n}(\cdot)$ is the Probability Density Function (PDF) of γ'_n. Likewise, we have: $-D_{w_n}\mathcal{O}_n > 0$.
- **Property A.3**: Without loss of generality, we can order the files according to their popularity. Based on this and property A.2, the values $\{q_n^*\}_1^N$ and $\{w_n^*\}_1^N$ are decreasing w.r.t. n.
- **Property A.4**: Based on A.1, there exists a $K \in S_n$ such that $w_n = q_n = 0$ for $n > K$.

- **Property A.5**: At the optimum solution of $P_{(\beta)}$, the constraints over cache capacity and total bandwidth are fulfilled with equality, i.e., $\sum_{n=1}^{N} w_n^* = 1$ and $\sum_{n=1}^{N} q_n^* = L$, based on A.2.

B. Solution Characteristic of Problem $P_{(\beta)}$

Formulating the Lagrange function for $P_{(\beta)}$, taking into account properties A.4 and A.5 and then taking the derivative with respect to w_n and q_n implies that there exists $K \in S_N$ such that:

$$v_1 = -f_n D_{w_n}\mathcal{O}(w_n, q_n), \quad v_2 = -f_n D_{q_n}\mathcal{O}(w_n, q_n), \quad (8)$$

for $n \leq K$, and $w_n = q_n = 0,$ for $n > K$, where v_1 and v_2 are the Lagrange multipliers related to the equality constraints. Based on Property A.2, v_1 and v_2 are positive. Further, according to Proposition 1, we get: $D_{\eta_n}\mathcal{O}_n = \frac{2}{\beta}\frac{q_n}{\eta_n} D_{q_n}\mathcal{O}_n$, so we have:

$$D_{w_n}\mathcal{O}_n = \frac{2\alpha\ \ln(2)}{\beta}\frac{q_n 2^{\alpha/w_n}}{w_n^2\left(2^{\alpha/w_n}-1\right)} D_{q_n}\mathcal{O}_n. \quad (9)$$

Let $v_0^2 = \dfrac{v_1}{v_2}$, then we get:

$$q_n = v_0^2 \frac{\frac{\beta}{2}\, w_n^2(2^{\alpha/w_n}-1)}{2^{\alpha/w_n}\ln(2)\alpha}, \text{ for } n \leq K. \quad (10)$$

By defining the function $h_0(w_n) = -D_{q_n}\mathcal{O}(w_n, q_n)$ with q_n being replaced based on (10), we obtain:

$$\frac{v_2}{f_n} = h_0(w_n), \text{ for } n \leq K. \quad (11)$$

Corollary 1. *The optimal solution of problem $P_{(\beta)}$ lies in the region where $D_{q_nq_n}\mathcal{O}_n\ D_{w_nw_n}\mathcal{O}_n - D^2_{w_nq_n}\mathcal{O}_n$ is positive. Moreover, for $\beta \geq 4$, the optimal solution is within the region where $P_{(\beta)}$ is convex.*

Proof: Based on Proposition 2, $P_{(\beta)}$ is convex if $D_{q_nq_n}\mathcal{O}_n\ D_{w_nw_n}\mathcal{O}_n - D^2_{w_nq_n}\mathcal{O}_n \geq 0$, for $\beta \geq 4$, and for all n for which $w_n > 0,\ q_n > 0$. This condition specifies a region w.r.t. q_n and w_n where we denote it by $R^*(w_n, q_n)$.

On the other hand, we pay attention to the curvature of $h_0(w_n)$, where it increases as a function of w_n until reaching a peak value and then decreases. Accordingly, solving (11) w.r.t. w_n yields two possible solutions; one in the increasing region of the function $h_0(.)$ and the other in the decreasing region. Based on property A.3. the optimal w_n is decreasing w.r.t. n. Due to this fact, considering (11) and that f_n is decreasing w.r.t. n imply that the optimal w_n should lie in the decreasing region of $h_0(w_n)$.

The decreasing region of $h_0(\cdot)$ is specified based on a condition w.r.t. w_n. However, in order to compare it with the convexity region $R^*(w_n, q_n)$, we need to express this condition w.r.t. (w_n, q_n). To do so, we differentiate $h_0(w_n)$ w.r.t. w_n and set it to zero. It leads to a condition w.r.t. w_n that describes the decreasing region. We then replace v_0^2 with $q_n \dfrac{2^{\alpha/w_n}\ln(2)\alpha}{\frac{\beta}{2}w_n^2(2^{\alpha/w_n}-1)}$, based on (10), to express this condition w.r.t. the pair (w_n, q_n). We denote this region by $R^0(w_n, q_n)$. From (8), we have:

$$v_0^2 = \frac{D_{w_n}\mathcal{O}(w_n, q_n, \beta)}{D_{q_n}\mathcal{O}(w_n, q_n, \beta)}, \text{ for } n \leq K, \quad (12)$$

which based on (10) can be represented as

$$v_0^2 = g_0(w_n) f_0(q_n), \quad (13)$$

where $g_0(w_n) = 2^{\alpha/w_n}/\big(\frac{\beta}{2}w_n^2(2^{\alpha/w_n}-1)\big)$ and $f_0(q_n) = \alpha\ln(2)q_n$. Equation (13) leads to $q_n = f_0^{-1}\big(v_0^2/g_0(w_n)\big)$. Now, according to (11) and (8), $h_0(\cdot)$ is equal to $A(w_n, q_n)$ with q_n replaced by $f_0^{-1}\big(v_0^2/g_0(w_n)\big)$, where $A(w_n, q_n) = A_n = -\frac{1}{f_n}D_{q_n}\mathcal{O}(w_n, q_n, \beta)$. To find $R^0(w_n, q_n)$ we differentiate $A(\cdot,\cdot)$ w.r.t. w_n. However, note that (12) holds for any q_n and w_n with $n \leq K$. It implies that v_0 does not depend on n. Therefore, using the chain rule and expressing the result in terms of (q_n, w_n), we get:

$$\begin{aligned}\frac{dA_n}{dw_n} &= -\frac{1}{f_n}\left(D_{w_nq_n}\mathcal{O}_n - D_{q_nq_n}\mathcal{O}_n \frac{v_0^2}{f_0'(q_n)}\frac{g_0'(w_n)}{g_0^2(w_n)}\right)\\ &= \frac{1}{f_n}D_{q_n}\mathcal{O}_n\ \frac{D_{q_nq_n}\mathcal{O}_n\ D_{w_nw_n}\mathcal{O}_n - D^2_{w_nq_n}\mathcal{O}_n}{D_{w_nq_n}\mathcal{O}_n\ D_{q_n}\mathcal{O}_n - D_{q_nq_n}\mathcal{O}_n\ D_{w_n}\mathcal{O}_n},\end{aligned}$$

where we computed $g_0'(w_n)/f_0'(q_n)$ from (12) to get the last line. Now, based on (9), we get $D_{w_nq_n}\mathcal{O}_n = \frac{2\alpha}{\beta}\frac{2^{\alpha/w_n}}{w_n^2\left(2^{\alpha/w_n}-1\right)}(D_{q_n}\mathcal{O}_n + q_n\ D_{q_nq_n}\mathcal{O}_n)$, which gives:

$$\begin{aligned}&D_{w_nq_n}\mathcal{O}_n\ D_{q_n}\mathcal{O}_n - D_{q_nq_n}\mathcal{O}_n\ D_{w_n}\mathcal{O}_n =\\ &\qquad\frac{2\alpha}{\beta}\frac{2^{\alpha/w_n}}{w_n^2\left(2^{\alpha/w_n}-1\right)} D^2_{q_n}\mathcal{O}_n \geq 0.\end{aligned}$$

Considering that $-D_{q_n}\mathcal{O}_n$ and the denominator of last line are positive, the region $R^0(w_n, q_n)$ is merely determined by the sign of $D_{q_nq_n}\mathcal{O}_n\ D_{w_nw_n}\mathcal{O}_n - D^2_{w_nq_n}\mathcal{O}_n$. In addition, for $\beta \geq 4$, based on Proposition 2, it confirms that $R^0(w_n, q_n) = R^*(w_n, q_n)$. ■

C. Generalizability of Solutions

We define the caching density per storage unit $p_n = q_n/L$ for $n \in S_N$ such that $\sum_{n=1}^{N} p_n = 1$ and $p_n \leq 1/L$. Hence, file n in the library is cached according to a PPP with intensity λ_{eff}, where $\lambda_{\text{eff}} = \lambda L$ is the effective intensity of a single-file cache. Accordingly, problem $P_{(\beta)}$ can be expressed based on $\{p_n\}_n$ and λ_{eff}. Note that the optimal solution of Problem $P_{(\beta)}$ leads to two types of solutions. In the first type, the solution is characterized by $p_n^* < 1/L$ for all $n \in S_N$. In the second type, there exists $p_n^* = 1/L$ for some $n \in S_N$. The former type of solution has the following property. If an optimal solution is obtained, it is still the optimal solution for a problem with the same λ_{eff} and α but with a cache capacity $L' > L$. Hence, for a problem of type $P_{(\beta)}$ with distinct $\tilde{\lambda}_{\text{eff}}$ and α, there exists a threshold L_{th} for which if problem $P_{(\beta)}$ is solved, its solution will be still optimum for any $L \leq L_{th}$. Note that L_{th} indicates the *maximum allowed cache capacity* for a specific setting of problem $P_{(\beta)}$. In fact, L_{th} will be the largest number for which $\max(\{p_n^*\}_1^N) \leq 1/L_{th}$. This property indicates that the optimal cache solution depends on the product of cache intensity λ and cache L capacity, and not both separately.

Furthermore, if the optimal solution of problem $P_{(\beta)}$ for a distinct N is obtained such that there exist $k_1 < N$ for which $p_n = w_n = 0$ for all $n > k_1$, then the solution is still optimum for the problem for $N' > N$ with the same skewness θ. Moreover, if $\mathcal{O}_{\text{tot}}$ denotes the overall outage probability with N files, we have,

$$\mathcal{O}'_{\text{tot}} = \frac{\sum_{n=1}^{N} n^{-\theta}}{\sum_{n=1}^{N'} n^{-\theta}} \mathcal{O}_{\text{tot}} + \frac{\sum_{n=N+1}^{N'} n^{-\theta}}{\sum_{n=1}^{N'} n^{-\theta}},$$

where $\mathcal{O}'_{\text{tot}}$ is the overall outage probability with N' files.

The two above properties discuss the generalizing properties of problem $P_{(\beta)}$. Given the solution of a specific problem and some conditions, the optimal solution to other caching problems can be obtained without the need to solve the problem again.

D. Optimal Solution Approach to $P_{(\beta)}$

The optimal caching solution of $P_{(\beta)}$ can be obtained based on the solution of $P_{(4)}$. Corollary 1 states that for any value of β, a solution of the Karush–Kuhn–Tucker (KKT) conditions of problem $P_{(\beta)}$ in the region where $D_{q_n q_n}\mathcal{O}_n \; D_{w_n w_n}\mathcal{O}_n - D^2_{w_n q_n}\mathcal{O}_n > 0$ for $n \in S_K$, provides the global optimum. However, it is not easy to solve the KKT conditions for an arbitrary value of β. Instead, we follow another approach.

First, we consider $\beta = 4$ for which we have a closed form outage probability. The optimal cache solution for $\beta = 4$ can then be found by analyzing the KKT conditions.

For other values of β, the caching problem $P_{(\beta)}$ is considered as a parametric optimization problem with continuous parameter β. We use a path-following method [33], to obtain the optimal cache solution for any value of β based on the optimal solution of $\beta = 4$. Note that based on Corollary 1, and [33], this strategy enables us to find the global solution of cache policy for any value of β. This shows the importance of finding the cache policy for $\beta = 4$ as a starting point to obtain the optimal cache solution for an arbitrary value of β.

E. Optimal Solution Approach to $P_{(4)}$

Thus, we concentrate on $\beta = 4$. In addition to being a starting point for finding generic solutions, the solution for $\beta = 4$ has interest in its own right, as this is a typical path-loss exponent for moderate and long distances [34]. We get:

Corollary 2. *Consider the OMPMC scheme described in Proposition 1, with path-loss exponent $\beta = 4$. Then, the overall outage probability is*

$$\mathcal{O}_{\text{tot}} = \sum_{n=1}^{N} f_n \operatorname{erfc}\left(\frac{\pi^2 \lambda q_n}{4\sqrt{\eta_n(4)}}\right), \tag{14}$$

where $\operatorname{erfc}(\cdot)$ *is the complementary error function.*

Proof: see [25]. ■

As a result, a caching optimization problem for $\beta = 4$ can be formulated based on (7), which is denoted by $P_{(4)}$ and $\mathcal{O}(w_n, q_n) = \mathcal{O}_n = \operatorname{erfc}\left(\frac{\tilde{\lambda} q_n}{\sqrt{2^{\alpha/w_n} - 1}}\right)$, where $\tilde{\lambda} = \frac{\pi^2\lambda}{4}\sqrt{\gamma_{\text{ref}}(4)}$.

In the sequel, we intend to devise a method to find the optimal solution of problem $P_{(4)}$. Note that in this case, we have:

$$h_0(w_n) = \frac{2\tilde{\lambda}}{\sqrt{\pi}\sqrt{2^{\frac{\alpha}{w_n}} - 1}} \exp\left(-\tilde{\lambda}^2 v_0^4 \left(2^{\frac{\alpha}{w_n}} - 1\right)\left(\frac{2w_n^2}{2^{\frac{\alpha}{w_n}} \alpha \ln(2)}\right)^2\right).$$

Hence, we need to find v_0, v_2 and the optimal solutions $\{w_n^*\}_1^{K^*}$ and $\{q_n^*\}_1^{K^*}$ where K^* represents the optimum value of K or the optimum number of cached files. However, to escape from the exhaustive search over these unknown variables, we limit the search space by exploiting the results obtained from the KKT analysis. In this respect, the following points are considered.

- The value of v_0 is not known in advance, and needs to be searched. This is done using a line search approach and is expedited by an upper-bound over v_0 denoted by $v_{0,\max}$.
- By applying Jensen's inequality over $1/h_0(w_n)$, which is a convex function w.r.t. w_n, we can obtain an upper-bound for v_2 as a function of v_0. In addition, considering that $w_n \leq 1$

Algorithm 1 The Solution of $P_{(4)}$.

Inputs: $\tilde{\lambda}$, $\{f_n\}_1^N$, L, γ_{ref}, α.
Outputs: Optimal solutions K^*, $\big(\{w_n^*\}_1^{K^*}, \{q_n^*\}_1^{K^*}\big)$ and $\mathcal{O}^*$.
for $K \in S_N$ **do**
  for $v_0 \in [0, v_{0,\max}]$ **do**
    for $v_2 \in [v_{2,\min}(v_0), v_{2,\max}(v_0)]$ **do**
      Apply a root-finding method over (11) to find $\{w_n\}_1^K$ so that $w_{\min} \leq w_n \leq 1$ and $\sum_{n=1}^K w_n = 1$.
    end for
    Compute q_n for $n \leq K$, based on (10).
  end for
  Select $\{q_n\}_1^K$ that minimize $|\sum_{n=1}^K q_n - L|^2$.
  Compute and stack $\mathcal{O} = \sum_{n=K+1}^N f_n + \sum_{n=1}^K f_n \operatorname{erfc}\Big(\frac{\tilde{\lambda} q_n}{\sqrt{2^{\alpha/w_n}-1}}\Big)$ based on selected $\{q_n\}_1^K$ and $\{w_n\}_1^K$.
end for
Select the minimum of stacked outage probability and its corresponding solution. Nominate the optimal solutions by $\mathcal{O}^*$, K^* and $\big(\{w_n^*\}_1^{K^*}, \{q_n^*\}_1^{K^*}\big)$.

for all $n \leq K$, leads to a lower-bound for v_2. Therefore, we have $v_{2,\min}(v_0) \leq v_2 \leq v_{2,\max}(v_0)$.

- The equivalence of regions $R^*(w_n, q_n)$ and $R^0(w_n, q_n)$, based on Corollary 1, dictates that $\min\big(\{w_n^*\}_1^K\big) \geq w_{\min}$, where $w_{\min}$ is the extremum of the function $h_0(w_n)$. It facilitates the process of finding a set $\{w_n\}_1^K$ that satisfies (11) for given v_2 and v_0.

The pseudo-code for finding the optimal solution of problem $P_{(4)}$ is presented in Algorithm 1. Since K^* and v_0 are not known in advance, we need to search for them in the corresponding limited intervals. We also seek for v_2 but limit the search space exploiting the fact that $v_{2,\min}(v_0) \leq v_2 \leq v_{2,\max}(v_0)$. We leverage a root-finding method for (11) to find $\{w_n\}_1^K$ such that the constraints $\sum_{n=1}^K w_n = 1$ is satisfied. We then find $\{q_n\}_1^K$ based on the obtained $\{w_n\}_1^K$ and using (10) so that $|\sum_{n=1}^K q_n - L|^2$ is minimized. Accordingly, we compute the overall outage probability and stack it for the current sought K. After we finalize the searching over possible candidates of K^*, we choose the minimum of the stacked outage probability and nominate it as the optimal solution.

Finding the optimal solution of problem $P_{(4)}$ needs to search over a 2-dimensional space for each value of $K \in S_N$, so it is time-consuming especially when K^* is large. This motivates us to consider a low-complexity sub-optimal solution as discussed in the sequel.

F. Low-complexity Sub-Optimal Solution

We devise a low-complexity solution for problem $P_{(4)}$, by considering a variable transformation $\xi_n := 1/(2^{\alpha/w_n} - 1)$, and utilizing an upper-bound of the resource allocation constraint in terms of the transformed variable, i.e., $\sum_{n=1}^N w_n = \sum_{n=1}^N \alpha/\log_2(1+\frac{1}{\xi_n}) = 1$. Therefore, problem $P_{(4)}$ is transformed into a new problem w.r.t. variables $\{\xi_n\}_1^N$ and $\{q_n\}_1^N$, with the objective $\sum_{n=1}^N f_n \mathcal{O}(\xi_n, q_n)$ and the transformed resource constraint $\sum_{n=1}^N \alpha/\log_2(1+\frac{1}{\xi_n}) = 1$, where $\mathcal{O}(\xi_n, q_n) = \operatorname{erfc}\big(\tilde{\lambda} q_n \sqrt{\xi_n}\big) \equiv \mathcal{O}\big(w_n(\xi_n), q_n\big)$. Based on property A.4 and the mentioned variable transformation, there exists $K \in S_N$ such that $\xi_n = 0$ for $n > K$. The resource constraint thus can be bounded as:

$$\sum_{n=1}^K \frac{\alpha}{\log_2(1+\frac{1}{\xi_n})} \leq \frac{K}{\log_2\Big(1+\frac{K}{\sum_{n=1}^N \xi_n}\Big)}, \tag{15}$$

which is obtained by applying Jensen's inequality and using the concavity of $1/\log_2(1+1/\xi)$ w.r.t. ξ. Therefore, the constraint $\sum_{n=1}^N w_n = 1$ in problem $P_{(4)}$ can be relaxed by the upper-bound given in (15) as: $\sum_{n=1}^N \xi_n = K(2^{\alpha K}-1)^{-1}$. Consequently, the relaxed optimization problem is obtained as:

$$\widetilde{P}_{(4)}(K): \quad \min_{\{q_n\}_1^N, \{\xi_n\}_1^N} \sum_{n=1}^N f_n \mathcal{O}(\xi_n, q_n), \qquad \text{s.t.} \begin{cases} \sum_{n=1}^N \xi_n = K_\alpha, & \sum_{n=1}^N q_n = L, \\ 0 \leq \xi_n \leq \alpha_2, & \text{for} \ \ n \in S_N, \\ 0 \leq q_n \leq 1, & \text{for} \ \ n \in S_N. \end{cases}$$

where, $K_\alpha = K(2^{\alpha K}-1)^{-1}$ and $\alpha_2 = 1/(2^\alpha - 1)$.

Corollary 3. *Problem* $\widetilde{P}_{(4)}(K)$ *is convex, if* $6\tilde{\lambda}^2 q_n^2 \xi_n - 1 \geq 0$ *for all* $n \leq K$.

Proof: The constraints of problem $\widetilde{P}_{(4)}(K)$ are affine, implying that this problem is convex if its objective function is convex. Using the Hessian matrix of $\mathcal{O}(\xi_n, q_n)$ w.r.t. ξ_n and q_n, we compute the leading principal minors of orders 1 and 2:

$$D_1 = \frac{4\tilde{\lambda}^3 q_n \sqrt{\xi_n^3}}{\sqrt{\pi}} \exp\big(-\tilde{\lambda}^2 q_n^2 \xi_n\big),$$
$$D_2 = \frac{\tilde{\lambda}^2(6\tilde{\lambda}^2 q_n^2 \xi_n - 1)}{\pi \xi_n} \exp\big(-\tilde{\lambda}^2 q_n^2 \xi_n\big).$$

This implies that $\mathcal{O}(\xi_n, q_n)$ is convex if $6\tilde{\lambda}^2 q_n^2 \xi_n - 1 \geq 0$. Considering the objective as a positive weighted sum of $\mathcal{O}(\xi_n, q_n)$, the statement follows. ∎

The condition $6\tilde{\lambda}^2 q_n^2 \xi_n - 1 \geq 0$ implies that as the cache intensity λ increases, q_n and ξ_n should take smaller values in order for $\widetilde{P}_{(4)}(K)$ to be convex, and as λ decreases, they should take larger values.

Hereafter, we denote the optimal solution of $\widetilde{P}_{(4)}(K)$ by $\{\widetilde{\xi}_n\}_1^K$ and $\{\tilde{q}_n\}_1^K$ for $K \in S_N$.

Problem $\widetilde{P}_{(4)}(K)$ is solved w.r.t. K and we use $\tilde{K}$ to denote the optimum value of K. Considering that $\tilde{K}$ is not known in advance, we consider a candidate set, denoted by S_K, within which $\tilde{K}$ might lie. We solve problem $\widetilde{P}_{(4)}(K)$ for all $K \in S_K$ and stack the values of the corresponding overall outage probabilities. We select the minimum of the stacked overall outage probabilities and nominate it as the optimal solution. The solution approach can be represented by a two-step optimization problem as:

$$\widetilde{P}_{(4)}: \min_{K \in S_K} \quad \min_{\substack{\{q_n\}_1^N, \{\xi_n\}_1^N \\ \sum_{n=1}^N q_n = L, \\ \sum_{n=1}^N \xi_n = K_\alpha,}} \quad \sum_{n=1}^{N} f_n \mathcal{O}(\xi_n, q_n), \tag{16}$$

where in the outer step, we search over $K \in S_K$ to find $\tilde{K}$ and in the inner step, we find the optimal solution $\{\tilde{\xi}_n\}_1^K$ and $\{\tilde{q}_n\}_1^K$ for the current value of K. Salient properties of $\widetilde{P}_{(4)}(K)$ are:.

- **Property B.1**: $\mathcal{O}(\xi_n, q_n)$ is a monotonically decreasing function w.r.t. ξ_n and q_n.
- **Property B.2**: Based on the property B.1 and the fact that f_n is decreasing w.r.t. n, $\{\tilde{\xi}_n\}_1^N$ and $\{\tilde{q}_n\}_1^N$ are decreasing w.r.t. n.

Theorem 1. *The solution of the KKT conditions of problem $\widetilde{P}_{(4)}(K)$ is given by:*

$$q_n = \begin{cases} h_1^{-1}\Big(\frac{v_2}{2\tilde{\lambda} f_n}\sqrt{\frac{\pi L}{K_\alpha}}\Big), & n \le K \\ 0, & n > K \end{cases}, \qquad \xi_n = \frac{K_\alpha}{L} q_n,$$

where $h_1(q_n) = \sqrt{q_n}\exp\left(-\tilde{\lambda}^2 q_n^3 K_\alpha\right)$, and v_2 is a Lagrange multiplier that is a solution of

$$\sum_{n=1}^{K} h_1^{-1}\Big(\frac{v_2}{2\tilde{\lambda} f_n}\sqrt{\frac{\pi L}{K_\alpha}}\Big) = 1.$$

Proof: Proof is not given due to lack of space. ∎

Corollary 4. *The optimal solution of problem $\widetilde{P}_{(4)}(K)$ lies in the region where it is convex.*

Proof: Based on Corollary 3, $\widetilde{P}_{(4)}(K)$ is convex if $6\tilde{\lambda}^2 q_n^2 \xi_n - 1 \ge 0$, for all n for which $q_n > 0,\ \xi_n > 0$. This condition specifies a region $\tilde{R}(\xi_n, q_n)$ w.r.t. q_n and ξ_n. In general, there is no guarantee that the optimal solution of $\widetilde{P}_{(4)}(K)$ should lie in this region. However, for this problem, it can be proved following the lines of Corollary 1. Further, we show that $\tilde{R}(\xi_n, q_n)$ is equal to the decreasing region of $h_1(\cdot)$. For this, we differentiate $h_1(\cdot)$ w.r.t. q_n which gives $h_1'(q_n) = \frac{\exp\left(-\tilde{\lambda}^2 q_n^3 K_\alpha/L\right)}{2\sqrt{q_n}}\left(1 - 6q_n^3\tilde{\lambda}^2 K_\alpha/L\right)$, then we replace K_α/L with ξ_n/q_n that gives the condition $6\tilde{\lambda}^2 q_n^2 \xi_n - 1 \ge 0$. ∎

Algorithm 2 The Solution of $\widetilde{P}_{(4)}$.

Inputs: $\tilde{\lambda}$, $\{f_n\}_1^N$, L, γ_{ref}, α.
Outputs: Optimal solutions $\tilde{K}$, $\left(\{\tilde{\xi}\}_1^{\tilde{K}}, \{\tilde{q}_n\}_1^{\tilde{K}}\right)$ and $\tilde{\mathcal{O}}$.
for $K \in [1, K_{\max}]$ **do**
  Set $K_\alpha = \frac{K}{2^{\alpha K}-1}$.
  for $v_2 \in [v_{2,\min}, v_{2,\max}]$ **do**
    Apply a root-finding method for (17) to find $\{q_n\}_1^K$ so that $q_{\min} \le q_n \le 1$ and $\sum_{n=1}^K q_n = L$.
  end for
  Set $\xi_n = \frac{K_\alpha}{L} q_n$, for $n \le K$.
  Compute and stack $\mathcal{O} = \sum_{n=K+1}^N f_n + \sum_{n=1}^K f_n \,\mathrm{erfc}(\tilde{\lambda} q_n \sqrt{\xi_n})$ based on the obtained $\{q_n\}_1^K$ and $\{\xi_n\}_1^K$.
end for
Select the minimum of stacked outage probability and its corresponding solution. Nominate the optimal solutions by $\tilde{\mathcal{O}}$, $\tilde{K}$ and $\left(\{\tilde{\xi}\}_1^{\tilde{K}}, \{\tilde{q}_n\}_1^{\tilde{K}}\right)$.

Note that based on Corollaries 3 and 4, as λ grows the value of K increases considering that $\sum_{n=1}^N q_n$ and $\sum_{n=1}^N \xi_n$ are upper-bounded. Consequently, the total resource can be allocated to more files. For the excessively large λ, the value of K tends to N, and ξ_n and q_n can take the maximum possible values N_α/N and L/N, receptively, for $n \in S_N$, where $N_\alpha = N(2^{\alpha N}-1)^{-1}$.

Based on Theorem 1, there exists $K \in S_N$ such that for $n > K: q_n = \xi_n = 0$ and for $n \le K$ we have:

$$\frac{v_2}{2\tilde{\lambda} f_n}\sqrt{\frac{\pi L}{K_\alpha}} = h_1(q_n). \tag{17}$$

Applying Jensen's inequality over the convex function $1/h_1(.)$ and considering that $q_n \le 1$, a lower-bound and an upper-bound can be found for v_2 implying $v_{2,\min} \le v_2 \le v_{2,\max}$. In addition, an upper bound can be found for K as $K \le \min\left(K_{\max}, N\right)$, where $K_{\max}^2\left(2^{\alpha K_{\max}} - 1\right) \le 6\tilde{\lambda}^2 L^2$. Note that based on Corollary 4, we should have $q_n \ge q_{\min}$ for $n \le K$, where $q_{\min}$ is the solution of $h_1'(q_n) = 0$. Correspondingly, we have $\xi_n \ge \xi_{\min}$ for $n \le K$, where $\xi_{\min} = \frac{K_\alpha}{L} q_{\min}$. Based on these expositions, we devise a method to find the optimal solution of $\widetilde{P}_{(4)}$. Algorithm 2 presents the pseudo-code for the devised algorithm. Since the value of $\tilde{K}$ is unknown, we need to search in the provided interval $[1, K_{\max}]$. We also search for v_2 but in a limited range i.e., $v_{2,\min} \leq v_2 \leq v_{2,\max}$. We leverage a root-finding method for (17) to find $\{q_n\}_1^K$ such that the constraints $\sum_{n=1}^{K} q_n = 1$ is satisfied. According to Theorem 1 we then set $\xi_n = \frac{K_\alpha}{L} q_n$ for $n \leq K$, and compute the overall outage probability for the current K. The smallest outage proability in the candidate set of K values provides the optimal solution.

Algorithm 2 provides an approximate solution of problem $P_{(4)}$. Considering a limited one-dimensional search for v_2, this solution has a much less computational complexity compared to Algorithm 1. It motivated us to analyze the outage probability gap between the solutions of Algorithms 1 and 2.

G. Gap between Optimal and Sub-optimal Solutions

Now we address the relaxation gap, or the outage probability difference $\mathcal{G} = \tilde{\mathcal{O}} - \mathcal{O}^*$ between the solution of Algorithms 1 and 2. Here, $\tilde{\mathcal{O}}$ and $\mathcal{O}^*$ are evaluated with respective optimal numbers of cached files $\tilde{K}$ and K^*. Hereafter, we assume that $\tilde{K} = K^*$ which makes $\mathcal{G}$ bounded from above and provides an upper-bound for $\mathcal{G}$. The regions $R^*(w_n, q_n)$ and $\tilde{R}(\xi_n, q_n)$ can be evaluated using Corollaries 1 and 4. These indicate that $R^*(w_n, q_n) \subseteq \tilde{R}(w_n, q_n)$, where $\tilde{R}(w_n, q_n)$ is the convexity region related to problem $\widetilde{P}_{(4)}$, i.e., $\tilde{R}(\xi_n, q_n)$ with ξ replaced with $1/(2^{\alpha/w} - 1)$. With $\{w_n^*\}_1^{K^*}$ the optimal solution of problem $P_{(4)}$, it follows that $\min(\{w_n^*\}_1^{K^*}) \geq \tilde{w}_{\min}$, where

$$\tilde{w}_{\min} = \alpha / \log_2(1 + 1/\xi_{\min}), \tag{18}$$

and $\xi_{\min}$ is the minimum value that the solution of problem $\widetilde{P}_{(4)}$ can take for $K = K^*$. The optimal solution of problem $P_{(4)}$ lies in the region $\tilde{R}(w_n, q_n)$, but the optimal solution of $\widetilde{P}_{(4)}$ does not necessarily reside in $R^*(w_n, q_n)$. However, it can be easily verified whether the optimal solution of $\widetilde{P}_{(4)}$ is in $R^*(w_n, q_n)$. This can be done using the optimal solution of $\widetilde{P}_{(4)}$ given by Algorithm 1, and the regions $R^*(w_n, q_n)$ and $\tilde{R}(w_n, q_n)$ evaluated from Corollaries 1 and 4.

To analyze the gap, we extend the concept of relative convexity [35] to multi-variable functions. For this, let $f(\boldsymbol{y}) : \mathbb{R}^l \to \mathbb{R}$ and $\boldsymbol{x} = \boldsymbol{g}(\boldsymbol{y}) : \mathbb{R}^l \to \mathbb{R}^m$ be two multivariate functions, with $\boldsymbol{g} = [g_1, \ldots, g_m]^\top$. We call $f(\boldsymbol{y})$ convex w.r.t. $\boldsymbol{g}(\boldsymbol{y})$, denoted by $f \succ \boldsymbol{g}$, if and only if $\left(f \circ \boldsymbol{g}^{-1}\right)(\boldsymbol{x})$ is a convex function w.r.t. $\boldsymbol{x}$ [35]. We then have:

Lemma 1. *If $f \succ \boldsymbol{g}$, then:*

$$f(\boldsymbol{y}) - f(\boldsymbol{y}_0) \leq \left[\nabla_{\boldsymbol{y}} f^\top \odot \Delta_{\boldsymbol{y}} g_1, \ldots, \nabla_{\boldsymbol{y}} f^\top \odot \Delta_{\boldsymbol{y}} g_m\right]^\top \left(\boldsymbol{g}(\boldsymbol{y}) - \boldsymbol{g}(\boldsymbol{y}_0)\right), \tag{19}$$

where $\odot$ is the component-wise product and $\Delta_{\boldsymbol{y}} g_i$, for $i = 1, \ldots, m$, is a l-dimensional vector with elements

$$[\Delta_{\boldsymbol{y}} g_i]_k = \begin{cases} \dfrac{1}{\partial g_i / \partial y_k}, & \textit{for } \left[D_{x_i} \boldsymbol{g}^{-1}(\boldsymbol{x})\right]_k \neq 0 \\ 0, & \textit{otherwise} \end{cases},$$

for $k = 1, \ldots, l$.

Proof: refer to Appendix B. ■

Theorem 2. *Assume that $\{\tilde{\xi}_n, \tilde{w}_n\}_1^{K^*}$ is an optimal solution of the relaxed problem $\widetilde{P}_{(4)}$ with v_1 the Lagrange multiplier of the total resource constraint. The relaxation gap between this solution, and an optimal solution of the original problem $P_{(4)}$ is upper bounded by*

$$\mathcal{G} \leq U_1\left(v_1, r_{\text{dev}}, \tilde{\zeta}_1\right) + U_2\left(v_1, \delta, \{\tilde{\zeta}_n\}_1^{K^*}\right),$$

where the total resource deviation of the $\widetilde{P}_{(4)}$ solution is $r_{\text{dev}} = 1 - \sum_{n=1}^{N} \tilde{w}_n$, the difference between the smallest and smallest possible non-zero $\tilde{w}_n$ is $\delta = \tilde{w}_{K^*} - \tilde{w}_{\min}$, and

$$\begin{aligned} U_1 &= \frac{1}{\alpha} v_1 \ln(2)\, \tilde{\zeta}_1\, r_{\text{dev}}, \\ U_2 &= \frac{1}{\alpha} v_1 \ln(2)\, \delta \left(\tilde{\zeta}_1 K^* - \sum_{n=1}^{K^*} \tilde{\zeta}_n\right) \\ \tilde{\zeta}_n &= \tilde{\xi}_n (1 + \tilde{\xi}_n) \log_2^2(1 + 1/\tilde{\xi}_n). \end{aligned}$$

Proof: We use Lemma 1 with

$$f(q_n, \xi_n) = \operatorname{erfc}\left(\tilde{\lambda} q_n \sqrt{\xi_n}\right),$$

$$\boldsymbol{y} = \boldsymbol{g}(q_n, \xi_n) = \left[q_n, \frac{\alpha}{\log_2(1 + 1/\xi_n)}\right]^\top,$$

where $\boldsymbol{y} = (q_n, w_n)$. Hence,

$$\nabla_{\boldsymbol{y}} f(q_n, \xi_n) = -\frac{2\tilde{\lambda}}{\sqrt{\pi}} e^{-\tilde{\lambda}^2 q_n^2 \xi_n} \left[\sqrt{\xi_n}, \frac{q_n}{2\sqrt{\xi_n}}\right]^\top,$$

and

$$\begin{aligned} \Delta_{\boldsymbol{y}} g_1 &= [1,\ 0]^\top, \\ \Delta_{\boldsymbol{y}} g_2 &= \left[0,\ \frac{\ln(2)}{\alpha} \xi_n (1 + \xi_n) \log_2^2(1 + \frac{1}{\xi_n})\right]^\top. \end{aligned}$$

Taking into account that $R^*(w_n, q_n) = R^0(w_n, q_n)$, we find that (q_n^*, w_n^*) lies in the convex region of $\operatorname{erfc}\left(\frac{\tilde{\lambda} q_n}{\sqrt{2^{\alpha/w_n} - 1}}\right)$ and equivalently (q_n^*, ξ_n^*) lies in the region in which $f(q_n, \xi_n) = \operatorname{erfc}\left(\tilde{\lambda} q_n \sqrt{\xi_n}\right)$ is convex w.r.t. $\boldsymbol{g}(q_n, \xi_n)$, where

$\xi_n^* = 1/(2^{\alpha/w_n^*} - 1)$. Therefore, based on Lemma 1 and assuming $R^*(w_n, q_n) \subseteq \tilde{R}(w_n, q_n)$ we have:

$$\begin{aligned}\mathcal{G} \leq & \sum_{n=1}^{K^*} f_n \operatorname{erfc}\left(\tilde{\lambda}\tilde{q}_n\sqrt{\tilde{\xi}_n}\right) - \sum_{n=1}^{K^*} f_n \operatorname{erfc}\left(\tilde{\lambda}q_n^*\sqrt{\xi_n^*}\right) \\ \stackrel{(a)}{\leq} & -\underbrace{\sum_{n=1}^{K^*} v_2(\tilde{q}_n - q_n^*)}_{=\,0} + \sum_{n=1}^{K^*} \frac{1}{\alpha} v_1 \ln(2)\tilde{\xi}_n(1+\tilde{\xi}_n)\times \\ & \log_2^2(1+1/\tilde{\xi}_n)\left(\frac{\alpha}{\log_2(1+1/\xi_n^*)} - \frac{\alpha}{\log_2(1+1/\tilde{\xi}_n)}\right),\end{aligned}$$

where (a) follows from Lemma 1. Note that the term $\alpha/\log_2(1+1/\xi_n^*) - \alpha/\log_2(1+1/\tilde{\xi}_n)$ can be positive or negative for different value of n, so we add and subtract the minimum value of this term (denoted by δ') to simplify the above equation, where

$$\delta' = -\min\left(\left\{\frac{\alpha}{\log_2(1+1/\xi_n^*)} - \frac{\alpha}{\log_2(1+1/\tilde{\xi}_n)}\right\}_1^{K^*}\right).$$

Taking into account that $\tilde{\xi}_n \leq \tilde{\xi}_1$ for all $n \in S_K$, and the function $\xi(1+\xi)\log_2(1+1/\xi)^2$ is monotonically increasing w.r.t. $\xi \in [0,1]$, we get:

$$\mathcal{G} \;\leq\; \frac{1}{\alpha} v_1 \ln(2)\left(\tilde{\zeta}_1 r_{\text{dev}} + \delta'\left(\tilde{\zeta}_1 K^* - \sum_{n=1}^{K^*}\tilde{\zeta}_n\right)\right),$$

where we used: $\sum_{n=1}^{K^*} \frac{\alpha}{\log_2(1+1/\xi_n^*)} - \sum_{n=1}^{K^*} \frac{\alpha}{\log_2(1+1/\tilde{\xi}_n)} = 1 - \sum_{n=1}^{K^*} \tilde{w}_n$. Considering that $\delta' = \tilde{w}_{K^*} - w_{K^*}^* \leq \delta$, where $\delta = \tilde{w}_{K^*} - \tilde{w}_{\min}$, the statement follows. ■

It is noteworthy that the solution of problem $P_{(4)}$ is not needed to compute the upper-bound of the outage probability gap. The term U_1 depends multiplicatively on r_{dev}, i.e., on how tightly the total resource constraint is satisfied through the solution of problem $\widetilde{P}_{(4)}$. As the value of r_{dev} decreases, U_1 decreases and consequently a better bound can be achieved. However, even if we reach a solution for which $r_{\text{dev}} = 0$, the term U_2 still remains, meaning that the performance gap between $P_{(4)}$ and $\widetilde{P}_{(4)}$ is inevitable.

H. *Complexity Analysis of Algorithm 1 and 2*

We evaluate the computational complexity of Algorithms 1 and 2, assuming the optimal cache solutions are desired with d-digit precision. Without loss of generality, we evaluate the complexity of these algorithms assuming the optimum value of K is given. The complexity order for Newton-Raphson root-finding algorithm of the function $f(x) = 0$, follows $O(C(d)\log(d))$, with $C(d)$ being the computational complexity needed to compute $f(x)/f'(x)$ with d-digit precision [36].

According to lines 4 and 5 of Algorithm 1, we need two line-searches over the intervals $[0, v_{0,max}]$ and $[v_{2,min}, v_{2,max}]$, for which we consider M_{v_0} and M_{v_2} points to be evaluated, respectively. By evaluating the form of $h_0(w_n)/h_0'(w_n)$, five terms containing $2^{c_0\alpha/w_n}$ appear, with c_0 being a constant, which are the most computational consuming ones. Hence, its complexity follows $C(d) = O\big(5d\log(d)^2\big)$, with regards to the computational complexity needed to evaluate exponential functions [36]. Taking into account that the root-finding algorithm should be applied for each w_n, $n \in S_K$, the complexity order of this algorithm approximately is $O\big(5M_{v_2}M_{v_0}Kd\log(d)^3\big)$. The computational complexity of line 10 can be ignored compared to the other parts of the algorithm.

On the other hand, based on line 5 of Algorithm 2, we need a line-search over the interval $[v_{2,min}, v_{2,max}]$, for which we consider M_{v_2} points to be evaluated. Moreover, $h_1(q_n)/h_1'(q_n) = \frac{1}{2}q_n(1 - 6q_n^3\tilde{\lambda}_{\text{eff}}K_\alpha/L)$. Hence, the computational complexity to evaluate $h_1(q_n)/h_1'(q_n)$ is $C(d) = O\big(3d\log(d)\big)$, considering the multiplication complexity with d-digit precision [37]. Taking into account that the root-finding should be applied for each $q_n, n \in S_K$, the complexity order of this algorithm approximately follows $O\big(3M_{v_2}Kd\log(d)^2\big)$.

It is noteworthy that the computational complexity of Algorithm 1 is approximately $M_{v_0}\log(d)$ times the complexity of Algorithm 2.

V. Other Transmission Policies

To benchmark the benefits of multipoint multicast caching, we consider heuristic approaches and a MPMC scheme without orthogonal resource allocation. We compare the outage performance of Algoritm 2 to these, and to Single-Point Multicast and Single-Point Unicast policies inspired by the literature, as well as a SPMC policy with NWW resource allocation.

A. *Heuristic OMPMC Approaches*

1) Uniform OMPMC Scheme: We consider a *uniform* cache policy where the caches and resources are identically allocated to all files. For this, we set: $w_n = q_n/L = 1/N$, for $n \in S_N$. We call the resulting cache policy $\texttt{OMP}-\texttt{u}$.

2) Most-popular OMPMC Scheme: A *popularity-related* cache policy is considered where the caches and resources are allocated to the L most popular files. As such, we set: $w_n =$

$q_n/L = 1/L$, for $1 \leq n \leq L$, and $q_n = w_n = 0$, otherwise. Both bandwidth and cache allocation thus follows an on/off principle. We call the resulting cache policy $\mathtt{OMP-m}$.

3) Threshold-based OMPMC Scheme: A simplified *threshold based* policy can be developed based on a threshold $\nu \in S_N$. For files less popular than file ν, no resources are allocated, while equal resources will be allocated to the other files. Therefore, if $n > \nu$, we set: $w_n = q_n = 0$, otherwise, we set $w_n = q_n/L = 1/\nu$. Here, bandwidth and cache allocation thus follows an on/off principle. As each BS can cache L files, we have $\nu \geq L$. The optimization problem thus becomes a discrete line search over the objective $F_\nu \operatorname{erfc}\left(\frac{\tilde{\lambda}L}{v\sqrt{\eta_\nu(\beta)}}\right) + 1 - F_\nu$, where $F_\nu = \sum_{n=1}^{\nu} f_n$, and $\eta_\nu(\beta) = \left(2^{\alpha\nu} - 1\right)/\gamma_{\mathrm{ref}}(\beta)$. Note that we name the resulting cache policy as $\mathtt{OMP-th}$.

4) Only Bandwidth-optimized OMPMC Scheme: We also consider a sub-optimal cache policy for which only the bandwidth allocation is optimized and the caching placement is based on a simple method. As such, we follow the cache placement strategy: $q_n = Lf_n$, for $n \in S_N$. However, in order to respect $q_n \leq 1$, we bound any q_n whose value exceeds 1, and increase others such that $\sum_{n=1}^{N} q_n = L$. We call the resulting cache policy $\mathtt{OMP-w}$.

B. Non-Orthogonal MPMC

In an MPMC cache policy with non-orthogonal bandwidth allocation, all BSs that have a file multicast it to the requesting UEs. The same radio resources are used for all transmissions, thus interference is caused by the transmissions from other BSs caching other files. Each BS can transmit at most one file in each time period, therefore L = 1 is assumed in this scheme. We call this cache policy NOMP. For this policy we have:

Proposition 3. *Assume non-orthogonal MPMC delivery in a network with BSs distributed according to a PPP with intensity λ, and through an environment with path-loss exponent $\beta = 4$, and files $n = 1, \cdots, N$ from a library with popularity f_n being cached with probability q_n. Then, the overall outage probability is:*

$$\mathcal{O}_{\mathrm{tot}}^{\mathrm{NOMP}} = \frac{1}{2} + \frac{2}{\pi}\sum_{n=1}^{N} f_n \int_0^\infty \left\{\frac{1}{w}\sin\left(w^2\eta' - \tilde{\lambda}w\Big(q_n - \frac{1-q_n}{\left(2^\alpha - 1\right)^{-\frac{1}{2}}}\Big)\right)\times \exp\left(-\tilde{\lambda}w\Big(q_n + \frac{1-q_n}{\left(2^\alpha - 1\right)^{-\frac{1}{2}}}\Big)\right)\right\}dw,$$

where $\tilde{\lambda} = \frac{\pi^2\lambda}{2\sqrt{2}}$ and $\eta' = (2^\alpha - 1)/\gamma_{\mathrm{ref}}(4)$. The optimization problem for this cache policy can be correspondingly formulated. The solution of this problem gives the optimal cache placement through MPMC transmission scheme without bandwidth allocation.

C. Single-point Multicast

Following [6], [10], [11], [15], we also consider a caching policy based on single-point multicast. For this, each UE is served by its nearest caching BS. Each BS acts with frequency reuse 1, i.e., no NWW resource allocation across files is applied. Hence, to respond to any file request, all other BSs, except the nearest BS caching the file of interest, interfere with the desired signal. To simplify analysis we assume that a BS dedicates the whole bandwidth to transmit a cached file, thus only one file can be stored at the cache of each BS implying $L = 1$. We denote this policy by Non-Orthogonal Single-Point (NOSP). Based on [38], the overall outage probability for this policy is:

$$\mathcal{O}_{\mathrm{tot}}^{\mathrm{NOSP}} = 1 - \sqrt{\frac{\pi^3\lambda^2}{4\eta'}}\sum_{n=1}^{N} f_n q_n \exp\left(\frac{\chi_n^2}{4\eta'}\right)\operatorname{erfc}\left(\frac{\chi_n}{2\sqrt{\eta'}}\right),$$

where $\chi_n = \pi\lambda q_n\Big(\sqrt{\eta'}\tan^{-1}\left(\sqrt{\eta'}\right) + \sqrt{\eta'}\frac{\pi(1-q_n)}{2q_n} + 1\Big)$, with $\tan^{-1}(\cdot)$ being the inverse tangent function.

In order to mitigate interference, it is natural to also consider a SPMC scheme with NWW resource allocation across files. We denote this policy by Orthogonal Single-Point (OSP). For this policy, each BS can simultaneously transmit several files in disjoint radio resources, so L can be greater than one. Based on the analytical expression in [27], the overall outage probability for this cache policy is:

$$\mathcal{O}_{\mathrm{tot}}^{\mathrm{OSP}} = 1 - \sum_{n=1}^{N}\frac{\sqrt{\pi^3}\lambda f_n q_n}{\sqrt{4\eta_n(4)}}\exp\left(\frac{\kappa_n^2}{\eta_n(4)}\right)\operatorname{erfc}\left(\frac{\kappa_n}{\sqrt{\eta_n(4)}}\right),$$

where

$$\kappa_n = \frac{\lambda\pi q_n}{2}\Big(1 + \sqrt{\eta_n(4)\gamma_{\mathrm{ref}}(4)}\tan^{-1}(\sqrt{\eta_n(4)\gamma_{\mathrm{ref}}(4)})\Big),$$

with $\eta_n(4)$ from (4).

Accordingly, for OSP and NOSP policies, the cache design problem can be formulated and the optimal solution can be obtained numerically.

D. On-demand Single-point Unicast

According to [7]–[12], we consider an on-demand delivery policy based on the Single-Point Unicast transmission scheme.

The SPUC scheme operates as a conventional cellular network, where a requesting UEs is served by its nearest BS. The requested file is then fetched from the core network and unicast to the UE. The performance of this scheme depends on the UE intensity λ_u, in contrast to the `OMP`, `OSP` and `NOSP` policies. We denote this policy by `SPU`. Based on [39], the outage probability for this policy is:

$$\mathcal{O}^{\text{SPU}} = c(\kappa,\rho)\sum_{u=1}^{\infty}\frac{\Gamma(u,\eta^{\text{SU}})}{B(u,\kappa-1)}\left(\frac{1}{1+\kappa\rho}\right)^{u-1},$$

where $\kappa = 3.575$, $\rho = \frac{\lambda}{\lambda_u}$, $c(\kappa,\rho) = \frac{1}{\kappa-1}\left(\frac{\kappa}{\kappa+1/\rho}\right)^{\kappa}$, $\eta^{\text{SU}} = 2^{\alpha}-1$, $B(\cdot,\cdot)$ is the beta function and

$$\Gamma(u,\eta^{\text{SU}}) = 1 - \frac{P_u}{{}_2F_1(-\frac{1}{2},1,\frac{1}{2},-\eta^{\text{SU}})},$$

with $P_u = \min(\frac{1}{u},1)$ and ${}_2F_1(\cdot)$ being the hypergeometric function. Note that the outage probability of SPUC scheme $\mathcal{O}^{\text{SPU}}$ depends on $\frac{\lambda}{\lambda_u}$, and not on both separately.

VI. Simulation Results and Discussion

A. Simulation Model

The performance of cache policies are compared with the following overall settings. There is a library consisting N files with Zipf popularity distribution. Caches with storage capacity L are distributed according to a PPP with different cache intensity values λ. We apply an Urban NLOS scenario from 3GPP standard [40] with carrier frequency 2 GHz and the transmission power of BSs 23 dBm. The antenna gain at the UE and BS are 0 dBi and 8 dBi, respectively, the noise-figure of UE is 9 dB, the noise spectrum density is -174 dBm and the bandwidth is 20 MHz. In this scenario, we consider different values of the path-loss exponent: $\beta \in [3,5]$. In a scenario with $\beta = 4$, the reference SNR becomes $\gamma_{\text{ref}}(4) = 0.105$. Since the reference distance is $x_{\text{ref}} = 1$ km, the caching density is computed in units of caches/km^2.

B. Comparing Algorithms 1 and 2

First we compare the exact Algorithm 1 and the relaxed Algorithm 2 in terms of time consumption and caching performance. Table I shows the elapsed times needed for Algorithm 1 and 2 to achieve the caching solution and also the corresponding solutions for a cellular network with the settings tabulated in the first column. Recalling the complexity analysis of Algorithms 1 and 2, analyzed in section IV-H, we need to set $M_{v_0} = 1000$ and $M_{v_2} = 200$ for the evaluation of corresponding intervals. These algorithms have been performed using Matlab R2021a on a 8×1.70 GHz Intel Core i5-10310U Processor, equipped with 16 GB of memory and 12 Mbytes of data cache. The computational complexity of Algorithm 1 is remarkably higher than Algorithm 2; as explained in section IV-H, Algorithm 1 is $M_{v_0}\log(d)$ times more complexity than Algorithm 2 for d-digit precision.

The optimum values of the overall outage probabilities are shown in the fourth and fifth columns for $P_{(4)}$ and $\widetilde{P}_{(4)}$, respectively. Based on the obtained optimal values, the outage probability gap is computed and shown in the sixth column. The upper-bound of the relaxation gap from Theorem 2 is in the seventh column. Further, the last column depicts how the total resource allocation constraint is satisfied by the solution of problem $\widetilde{P}_{(4)}$.

According to this table, the solution of problem $\widetilde{P}_{(4)}$ is close to the solution of problem $P_{(4)}$ as confirmed by the value of overall outage probability and also by the small value of gap between the two solutions. Moreover, for the case where the solution of problem $P_{(4)}$ is not accessible due to computational complexity, the upper-bound of relaxation gap verifies the competency of using the solution of $\widetilde{P}_{(4)}$. The total resource allocation constraint is properly satisfied, by the solution of problem $\widetilde{P}_{(4)}$ based on the last column. These results validate the merit of the solution obtained using Algorithm 2 and justify its utilization instead of solving problem $P_{(4)}$.

C. Performance Results

Now we evaluate the performance of OMPMC variants and compare them to the benchmark schemes of Section V. The evaluated cache policies and their properties are tabulated in Table II. The results obtained by Algorithm 2 are siply denoted by `OMP`.

First, in order to compare to NOSP and NOMP, we consider a scenario with $L = 1$. We have $N = 10$, Zipf skewness as $\theta = 2.6$ and the spectral threshold $\alpha = 1/4$. Higher storage capacity scenarios can be considered as a special case when the cache intensity is scaled while the storage capacity is kept at $L = 1$.

For this scenario, the overall outage probability is plotted in Figure 1 as a function of the cache intensity. The superiority of `OMP`, against other policies, is largely due to the multipoint delivery strategy. The denser the BSs are distributed, the more power a user can receive from the BSs to detect the requested file without interference. All considered methods

TABLE I: Comparison Between the Solution of $P_{(4)}$ and $\widetilde{P}_{(4)}$

Parameters					Elapsed time [s]		Outage probability		Relaxation gap		r_{dev}
N	L	λ_{eff}	θ	α	Alg. 1	Alg. 2	$P_{(4)}$	$\widetilde{P}_{(4)}$	Simul.	Bound, Thm. 2	
5	2	12	0.6	1	90	5	0.173	0.173	6.5×10^{-5}	1.9×10^{-3}	0.999
5	2	16	2.6	1	92	5	0.01568	0.0157	6.8×10^{-5}	6.1×10^{-4}	0.998
20	4	80	1.6	1/2	1430	31	0.0586	0.0587	2.2×10^{-5}	7.76×10^{-4}	0.999
20	4	80	2.6	1/3	2272	38	0.00316	0.00317	5.6×10^{-6}	1.18×10^{-4}	0.998
100	10	5×10^4	1.6	1/3	N/A	86	N/A	0.0195	N/A	4.9×10^{-4}	0.999
100	20	2×10^4	2.6	1/4	N/A	73	N/A	3.33×10^{-4}	N/A	1.21×10^{-5}	0.999

TABLE II: Cache Policy Approaches

Policy	NWW Bandwidth Allocation	Cache Allocation	Transmission Scheme
`OMP`	Yes	Yes	OMPMC
`OMP-w`	Yes	$p_n \sim f_n$	OMPMC
`OMP-th`	On-off-threshold	On-off-threshold	OMPMC
`OMP-m`	On-off-popularity	On-off-popularity	OMPMC
`OMP-u`	Uniform	Uniform	OMPMC
`NOMP`	No	Yes	MPMC
`OSP`	Yes	Yes	SPMC
`NOSP`	No	Yes	SPMC
`SPU`	No	No	SPUC

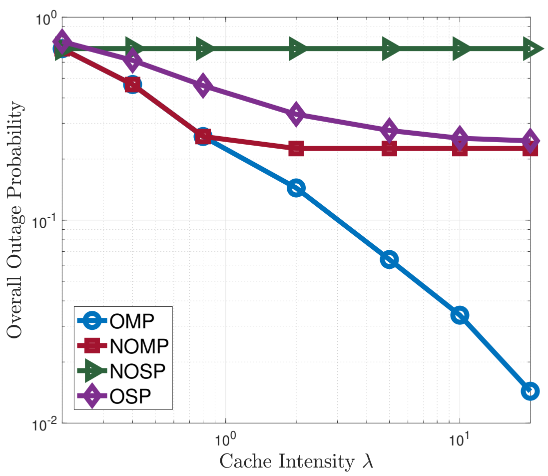

Fig. 1: Overall outage probability as a function of cache intensity for $\theta = 2.6$, $N = 10$ and $L = 1$.

except `OMP`, saturate when the cache intensity increases. Accordingly, it can be inferred that usage of non-orthogonal transmission in both single-point and multipoint multicast schemes is not an appropriate approach compared to the orthogonal transmission. Further, for the case the network characterizes files based on their popularity, employing the OMPMC transmission scheme is more recommendable than single-point delivery scheme.

However, we intend to evaluate this result for a larger content library. As a larger scenario we next consider settings based on [11]. The number of files increases to $N = 200$

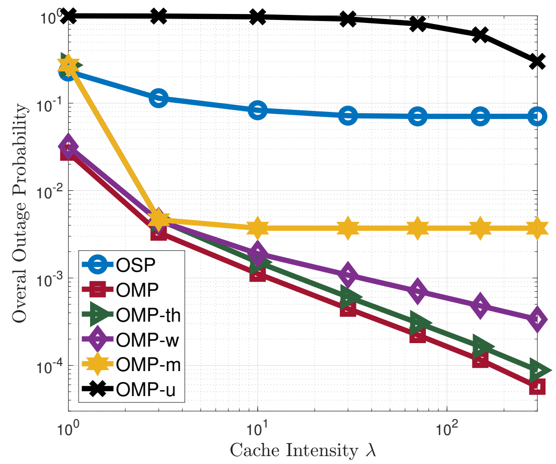

Fig. 2: Overall outage probability as a function of cache intensity for $\theta = 2.6$, $N = 200$ and $L = 20$.

and value of cache capacity to $L = 20$. Two values for the skewness are considered; $\theta = \{0.8, 2.6\}$. The spectral threshold is set to $\alpha = 0.05$. Notice that NOMP and NOSP are not compatible with this scenario, as $L > 1$.

The overall outage probability in this scenario is plotted in Figure 2 as a function of cache intensity for skewness $\theta = 2.6$. The `OMP` policy outperforms other cache strategies for all evaluated cache intensities. By increasing cache intensity, the effect of thermal noise is decreasing for `OSP` policy, and the network behaves in an interference-limited regime. The multicast multipoint delivery methods lead to a noise-limited regime, and the noise limitation decreases with increasing cache intensity, as the received power grows. As the intensity increases, $\mathtt{OMP-w}$ and $\mathtt{OMP-m}$ schemes diverge from the optimal solution. The $\mathtt{OMP-u}$ policy has the worst permanence as expected. For approximately all evaluated cache intensities, the $\mathtt{OMP-th}$ gives an acceptable result.

The overall outage probability is shown in Figure 3 for $\theta = 0.8$. The `OMP` and $\mathtt{OMP-th}$ policies outperform `OSP` in this scenario as well. As the cache intensities grows, the solution of $\mathtt{OMP-th}$ becomes more reliable compared to the result of $\mathtt{OMP-w}$.

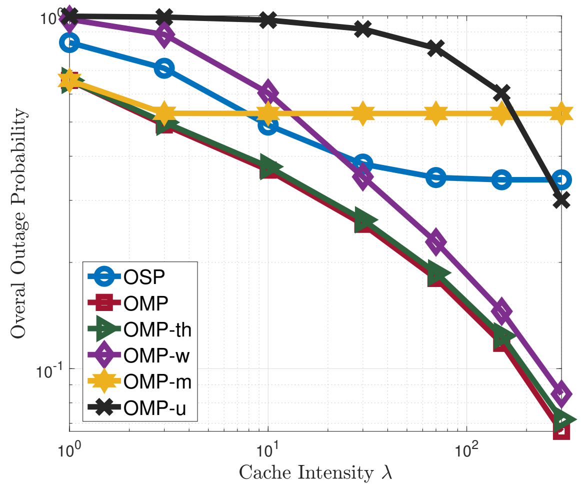

Fig. 3: Overall outage probability as a function of cache intensity for $\theta = 0.8$, $N = 200$ and $L = 20$.

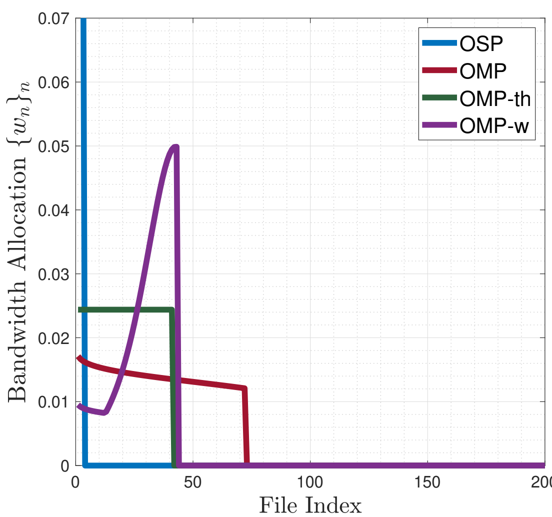

Fig. 4: Bandwidth allocation as a function of file index for $\lambda = 30$ and $\theta = 2.6$.

Figure 4 shows the resource allocation of cache solution as a function of the file index for cache intensity $\lambda = 30$ and skewness $\theta = 2.6$. For the $\texttt{OSP}$ policy, the total resource is allocated only for a few popular files, but for the $\texttt{OMP}$ policy around 40 percent of files can accommodate the total bandwidth. For the latter policy, the allocated resources decrease as specified based on property A.2, while for the $\texttt{OMP}-\texttt{w}$ they grow since the caching weights are not optimized.

According to the results of Figures 2 and 3, it can be inferred that employing the OMPMC delivery scheme has an affirmative effect on the network performance from outage probability and spectral efficiency perspective, especially for the intermediate and large values of the cache intensities. Therefore, for a cache-enabled network usage of the OMPMC delivery is more promising compared to the conventional

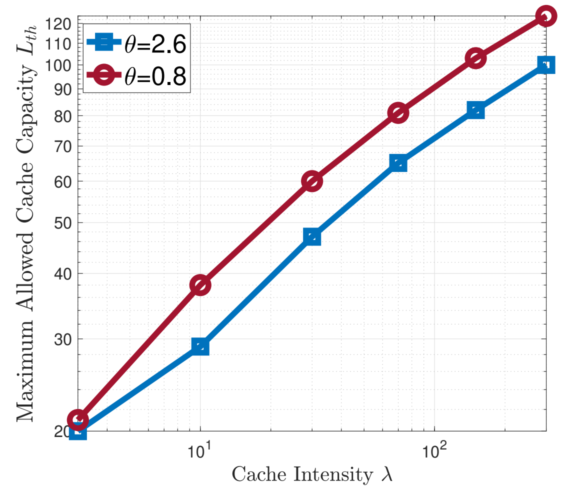

Fig. 5: Solution generalizabilty w.r.t. cache capacity as a function of cache intensity for $L = 20$.

single-point scheme. As a technical requirement, this advantage is obtained at the cost of the network being able to make the BSs synchronously cooperate during cache delivery phase. However, note that having a network with high-precisely synchronized BSs is feasible [30].

To investigates the solution generalizabilty of OMPMC policy w.r.t. cache capacity, the maximum allowed capacity L_{th} is obtained and plotted in Figure 5 for different values of cache intensities, (as presented in Section IV-C). Note that L_{th} indicates how much the cache capacity can increase without the cache solution being changed, for a given optimal cache policy of OMPMC scheme. For both values of skewness, as the intensity grows, the value of L_{th} increases, without the optimal solution being changed. The reason is that as λ_{eff} increases, the resource is more abundant and the ratio of cached files grows which leads to decrease in q_1. The reduction of q_1, in turn, causes L_{th} to increase. Figure 6 evaluates the solution generalizabilty of OMPMC policy w.r.t. number of files, for cache intensity $\lambda = 30$. The generalizable solutions are colored with blue for $\theta = 0.8$ and red for $\theta = 2.6$. According to Section IV-C, the minimum values of N, for which the solution generalizability holds, are $N = 81$ and $N = 71$ for vales of skewness 0.8 and 2.6, respectively. The optimal cache policy for any value of $N' > N$ can be obtained based on the solution of N, and the optimal outage probability can be found using equation given in section IV-C.

Although the cache policy is analyzed and obtained for path-loss exponent $\beta = 4$, as discussed in Section IV-D, the solution can be generalized for other values of β. Here, we

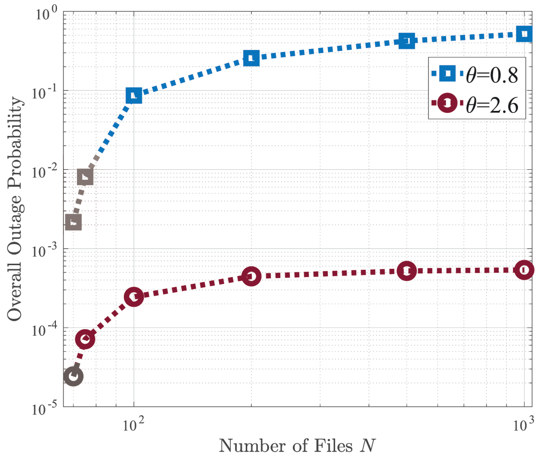

Fig. 6: Solution generalizabilty as a function of number of files for $\lambda = 30$ and $L = 20$.

consider different values for the path-loss exponent as $\beta \in [3, 5]$, and extract the optimal cache policy of Table II using a Path-Following (PF) method [33]. We call the optimal cache policy of this approach $\mathtt{Opt-PF}$. Apart from these policies, we also consider another strategy where the optimal solution of Algorithm 2 with $\beta = 4$ is obtained and then directly evaluated in the caching model (Proposition 1) with target β. We name the solution of this policy $\mathtt{Eval}$. Here, we use the legends $\mathtt{Th}$ and $\mathtt{W}$ to abbreviate the $\mathtt{OMP-th}$ and $\mathtt{OMP-w}$ policies.

Figure 7 shows the overall outage probability as a function of path-loss exponent for skewness $\theta = 2.6$ and cache intensity $\lambda = 30$. The $\mathtt{Opt-PF}$ cache policy outperforms other policies for all evaluated path-loss exponents. Moreover, $\mathtt{Th}$ and $\mathtt{W}$ give acceptable solutions, though $\mathtt{Th}$ outperforms $\mathtt{W}$. Further, $\mathtt{Eval}$ policy cannot give a reliable solution specially for large values β.

It is paramount to understand the relationship between network-wide multicasting and single-point unicast. For this, we compare the outage probabilities of these two schemes. When considering on-demand video delivery, SPUC has the benefit that the delivery can start whenever a user demands the file. If OMPMC delivery is applied to satisfy on-demand video delivery, the logical files managed by the delivery system should be small chunks of the videos [41], with each chunk having its own popularity. The first chunks of a video file would have the highest popularity, and would also be the most time critical for delivery to the user. Accordingly, for on-demand video delivery, OMPMC would operate on the level of video chunks.

For comparing OMPMC with SPUC, we thus consider a

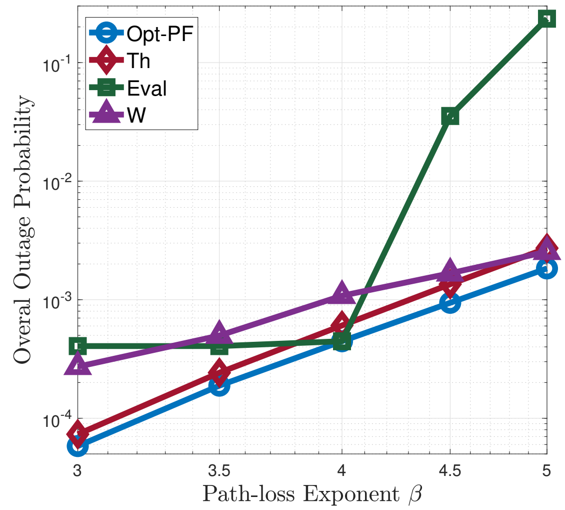

Fig. 7: Outage probability as a function of path-loss exponent for $\theta = 2.6$ and cache intensity $\lambda = 30$.

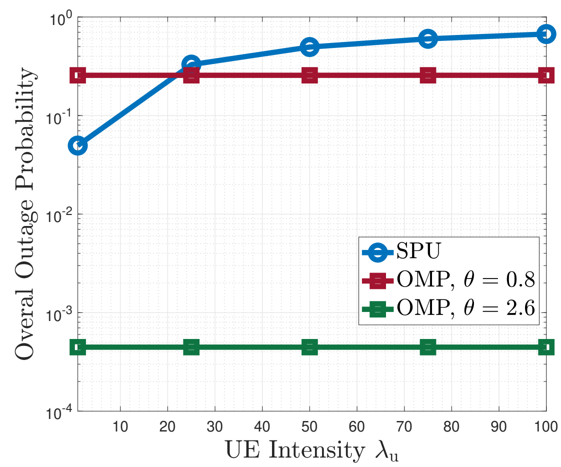

Fig. 8: Outage probability as a function of UE intensity λ_u and cache intensity $\lambda = 30$.

large library of $N = 1000$ file chunks, and otherwise take typical numbers from the scenarios above; $L = 20$, $\alpha = 0.05$ and $\theta \in \{0.8, 2.6\}$. The cache intensity is fixed at $\lambda = 30$, and the UE intensity λ_u changes from 1 to 100 UEs/km^2. Figure 8 shows the outage probability for OMPMC and on-demand SPUC schemes as a function of UE intensity λ_u. Note that the OMPMC policy is insensitive with respect to the UE intensity. Furthermore, SPUC treats all files separately and is thus independent of file popularity. For $\theta = 0.8$, there is cross-over behavior—at small UE intensity $\mathtt{SPU}$ outperforms $\mathtt{OMP}$, while when UE intensity increases this is reversed in a natural manner; the outage performance of SPUC depends on λ/λ_u. For a more skew popularity distribution with $\theta = 2.6$, $\mathtt{OMP}$ remarkably outperforms $\mathtt{SPU}$. The superiority of OMP against SPU is related to the multipoint delivery of OMPMC scheme and from the co-channel interference SPUC suffers from. It should be kept in mind, that the long tail of less popular files has to be served by an SPUC component of the network in any case.

VII. Conclusion

In this paper, we considered cache placement and delivery based on network-wide orthogonal multipoint multicast (OMPMC) transmission scheme. A cache policy jointly optimizing cache placement and radio resource allocation was formulated, based on the derived expression of the outage probability. An approach was developed to find the optimal policy for a generic path-loss environment, based on the optimal solution with path loss exponent $\beta = 4$. An algorithm was devised to obtain the optimal solution for $\beta = 4$, where the opitimization objective can be written in closed form. A low-complexity algorithm was proposed to obtain a sub-optimal solution. The computational complexity and relaxation gap between the optimal and sub-optimal solutions were evaluated and an upper-bound on the relaxation gap was obtained.

The solution is characterized by a file popularity threshold, with files less popular not being cached nor delivered at all. The outage probability depends only on a combination of cache probability and a channel gain threshold, and not on both separately. Furthermore, the optimal cache solution depends on the product of BS density and cache capacity, and not on both separately. These properties allow us to generalize solutions from one set of parameters to a continuum of other sets.

We compared the outage performance of OMPMC based on the devised low-complexity algorithm with single-point multicast, single-point unicast and other delivery approaches from the literature. Simulation results showed that the outage probability of OMPMC scheme outperforms other delivery strategies for different cache intensities, UE intensities and spectral efficiency thresholds. The usage of the OMPMC delivery scheme can be considered as a promising technique to be employed by the cache-enabled networks.

References

[1] U. Niesen, D. Shah, and G. Wornell, “Caching in wireless networks,” in *IEEE Int. Symp. Inf. Theory*, 2009, pp. 2111–2115.

[2] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] M. Choi, A. F. Molisch, D.-J. 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.

[9] 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.

[10] 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.

[11] 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.

[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] Z. Wang, Z. Cao, Y. Cui, and Y. Yang, “Joint and competitive caching designs in large-scale multi-tier wireless multicasting networks,” in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, 2017, pp. 1–7.

[14] 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.

[15] C. Ye, Y. Cui, Y. Yang, and R. Wang, “Optimal caching designs for perfect, imperfect, and unknown file popularity distributions in large-scale multi-tier wireless networks,” *IEEE Trans. Commun.*, vol. 67, no. 9, pp. 6612–6625, 2019.

[16] M. Bayat, R. K. Mungara, and G. Caire, “Achieving spatial scalability for coded caching via coded multipoint multicasting,” *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 227–240, Jan. 2019.

[17] X. Peng, Y. Shi, J. Zhang, and K. B. Letaief, “Layered group sparse beamforming for cache-enabled green wireless networks,” *IEEE Trans. Commun.*, vol. 65, no. 12, pp. 5589–5603, 2017.

[18] 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.

[19] 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.

[20] W. Sun, Y. Li, C. Hu, and M. Peng, “Joint optimization of cache placement and bandwidth allocation in heterogeneous networks,” *IEEE Access*, vol. 6, pp. 37 250–37 260, 2018.

[21] 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.

[22] 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.

[23] 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. (ICEEI)*, Jul. 2011, pp. 1–5.

[24] 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.

[25] M. Haenggi and R. Ganti, "Interference in large wireless networks," *Found. Trends Netw.*, vol. 3, pp. 127–248, 01 2009.

[26] 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.

[27] J. G. Andrews, A. K. Gupta, and H. S. Dhillon, "A primer on cellular network analysis using stochastic geometry," preprint ArXiv 1604.03183, Apr. 2016.

[28] M. Amidzade, H. Al-Tous, O. Tirkkonen, and G. Caire, "Cellular network caching based on multipoint multicast transmissions," in *Proc. IEEE Glob. Commun. Conf. (GLOBECOM)*, Dec. 2020, pp. 1–6.

[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] S. Ruffini, M. Johansson, B. Pohlman, and M. Sandgren, "5G synchronization requirements and solutions," *Ericsson Tech. Rev.*, pp. 1–13, Jan. 2021.

[31] M. Batariere, K. Baum, and T. P. Krauss, "Cyclic prefix length analysis for 4G OFDM systems," in *Proc. IEEE Veh. Technol. Conf. (VTC)*, Sep. 2004, pp. 543–547.

[32] E. Eriksen, "Lecture 5: Principal minors and the hessian," *BI Norwegian School of Management. Department of Economics.*, Oct. 2010.

[33] 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.

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

[35] T. Rajba, "On some relative convexities," *J. Math. Anal. Appl.*, vol. 411, no. 2, pp. 876 – 886, 2014.

[36] R. P. Brent, "Multiple-precision zero-finding methods and the complexity of elementary function evaluation," *Computing Research Repository*, vol. abs/1004.3412, May. 2010.

[37] D. Harvey and J. van der Hoeven, "Integer multiplication in time O(n log n)," *Annals of Mathematics*, vol. 193, no. 2, pp. 563–617, 2021.

[38] N. Garg, M. Sellathurai, B. Bettagere, V. Bhatia, and T. Ratnarajah, "Online learning models for content popularity prediction in wireless edge caching," in *Asilomar Conf. Signals, Syst., Comput.*, 2019, pp. 337–341.

[39] M. Amidzade, H. Al-Tous, O. Tirkkonen, and G. Caire, "Cellular traffic offloading with optimized compound single-point unicast and cache-based multipoint multicast," in *IEEE Wireless Commun. and Net. Conf. (WCNC)*, 2022, pp. 2268–2273.

[40] ETSI TR 125 942: Universal Mobile Telecommunications System (UMTS); RF system scenarios (3GPP TR 25.942 version 14.0.0), April 2017.

[41] L. Wang, S. Bayhan, and J. Kangasharju, "Optimal chunking and partial caching in information-centric networks," *Computer Communications*, vol. 61, pp. 48–57, 2015.

Appendix

A. *Proof of $D_{q_nq_n}\mathcal{O}_n > 0$ for $\beta \geq 4$.*

Based on 1, we have:

$$\begin{aligned}\mathcal{O}_n &= \frac{4}{\pi}\int_0^\infty \frac{1}{w}\cos\left(q_n\alpha_\beta w^{\frac{4}{\beta}}\right)\exp\left(-q_n\gamma_\beta w^{\frac{4}{\beta}}\right)\sin\left(w^2\eta_n\right)dw\\ &\overset{(a)}{=} \mathcal{L}^{-1}\left(\frac{1}{s}\exp\left(-q_n\lambda_\beta d_\beta\, s^{\frac{2}{\beta}}\right), \eta_n\right),\end{aligned} \tag{20}$$

where $\alpha_\beta = \lambda_\beta/\cos\left(\frac{\pi}{\beta}\right)$, $\gamma_\beta = \lambda_\beta/\sin\left(\frac{\pi}{\beta}\right)$, $\lambda_\beta = \dfrac{\pi^2\lambda}{\beta}$, $d_\beta = \tan(\pi/\beta) + \cot(\pi/\beta)$ and $\mathcal{L}^{-1}(\cdot, t)$ stands for the inverse Laplace transform w.r.t. s and evaluated at t. For (a), we use the contour integral $\oint_A \frac{z}{z^4+s^2}\exp\left(q_n(\alpha_\beta j - \gamma_\beta)z^{4/\beta}\right)dz$, $z \in \mathbb{C}$. Using the residue analysis and defining $F(z) = \frac{z}{z^4+s^2}\exp\left(q_n(\alpha_\beta j - \gamma_\beta)z^{4/\beta}\right)$, the integral can be computed. Accordingly, we have:

$$D_{q_nq_n}\mathcal{O}_n = \lambda_\beta^2 d_\beta^2\mathcal{L}^{-1}\left(s^{\frac{4}{\beta}-1}\exp\left(-q_n\lambda_\beta d_\beta\, s^{\frac{2}{\beta}}\right)\right).$$

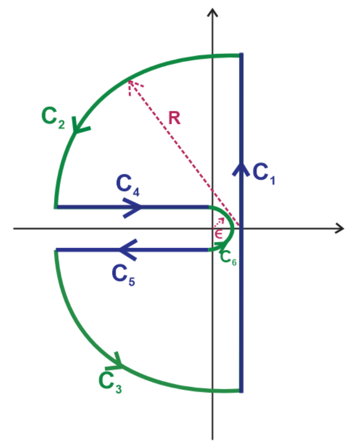

Fig. 9: Curve C, which is combination of the curves $\{C_i\}_{i=1}^6$. Note that there exists a branch point at origin.

To evaluate the above inverse Laplace transform, we take into account the complex integral $\oint_C z^{\frac{4}{\beta}-1}\exp\left(-\mu_\beta s^{\frac{2}{\beta}}\right)dz$ along curve C, where $\mu_\beta = q_n\lambda_\beta d_\beta$ and curve C is shown in Figure 9. Note that the function $F_2(z) = z^{\frac{4}{\beta}-1}\exp\left(-\mu_\beta z^{\frac{2}{\beta}}\right)$ does not have any residue inside the region covered by the curve C. As such, we have:

$$\oint_C F_2(z)e^{zt}dz = \sum_{i=1}^{6}\int_{C_i} F_2(z)e^{zt}dz = 0.$$

It can be shown that $\left|\int_{C_i} F(z)e^{zt}dz\right| \to 0$ for $i \in \{2, 3, 6\}$ when $R \to \infty$ and $\epsilon \to 0$. Moreover, based on Mellin's inverse formula, we get: $\int_{C_1} F_2(z)e^{zt}dz = 2\pi j\,\mathcal{L}^{-1}\big(F_2(s), t\big)$. We have:

$$\begin{aligned}2\pi j\,\mathcal{L}^{-1}\big(F_2(s), t\big) = &\int_{+\infty}^{0}(xe^{\pi j})^{\frac{4}{\beta}-1}e^{-\mu_\beta(xe^{\pi j})^{\frac{2}{\beta}}}e^{-xt}dx\\ &+\int_0^{+\infty}(xe^{-\pi j})^{\frac{4}{\beta}-1}e^{-\mu_\beta(xe^{-\pi j})^{\frac{2}{\beta}}}e^{-xt}dx.\end{aligned}$$

Hence, we obtain:

$$\begin{aligned}D_{q_nq_n}\mathcal{O}_n = \frac{\lambda_\beta^2 d_\beta^2}{\pi}\int_0^\infty x^{\frac{4}{\beta}-1}\exp\Big(-x\eta_n - \mu_\beta\cos\big(\frac{2\pi}{\beta}\big)x^{\frac{2}{\beta}}\Big)\\ \sin\Big(\frac{4\pi}{\beta} - \mu_\beta\sin\big(\frac{2\pi}{\beta}\big)x^{\frac{2}{\beta}}\Big)dx.\end{aligned}$$

For $\beta \geq 4$, this integral is positive considering the angle addition formula for sine and monotonically decreasing behavior of $\exp\left(-x\eta_n - \mu_\beta\cos(\frac{2\pi}{\beta})x^{\frac{2}{\beta}}\right)$ w.r.t. x.

B. *Proof of Lemma 1.*

Since $\left(f \circ \boldsymbol{g}^{-1}\right)(\boldsymbol{x})$ is a convex function, we can write

$$f\big(\boldsymbol{y}(\boldsymbol{x})\big) - f\big(\boldsymbol{y}(\boldsymbol{x}_0)\big) \leq \nabla_{\boldsymbol{x}}f^\top(\mathbf{x}-\mathbf{x}_0) = \nabla_{\mathbf{x}}f^\top\big(\mathbf{g}(\mathbf{y})-\mathbf{g}(\mathbf{y}_0)\big),$$

where $\mathbf{y}(\mathbf{x}) = \mathbf{g}^{-1}(\mathbf{x})$ and $\nabla_{\mathbf{x}} f = [\frac{df}{dx_1}, \ldots, \frac{df}{dx_m}]^\top$. Now, using the chain rule, we get:

$$\frac{df}{dx_i} = \sum_{j=1}^{l} \frac{\partial f}{\partial y_j}\frac{dy_j}{dx_i} = \sum_{j=1}^{l} \frac{\partial f}{\partial y_j} \cdot \left(\frac{1}{\frac{dx_i}{dy_j}}\right), \tag{21}$$

which obtained if $\frac{dx_i}{dy_j} \neq 0$. However, if y_j does not depend on x_i, it implies $\frac{dy_j}{dx_i} = 0$. Based on (21), we thus have:

$$\frac{df}{dx_i} = \nabla_{\mathbf{y}} f^\top \left((\nabla_{\mathbf{y}} g_i)_{\text{inv}}\right),$$

where subscript "inv" indicates component-wise inversion, so that for vector like $\mathbf{x} = [x_1, \ldots, x_l]^\top$, $\mathbf{x}_{\text{inv}} = [1/x_1, \ldots, 1/x_l]^\top$. Therefore, we can write:

$$\nabla_{\mathbf{x}} f = [\nabla_{\mathbf{y}} f^\top \odot (\nabla_{\mathbf{y}} g_1)_{\text{inv}}, \ldots, \nabla_{\mathbf{y}} f^\top \odot (\nabla_{\mathbf{y}} g_m)_{\text{inv}}]^\top.$$

By defining $\Delta_{\mathbf{y}} g_i = (\nabla_{\mathbf{y}} g_i)_{\text{inv}}$, Lemma 1 follows.

Mohsen Amidzadeh received the B.Sc. and M.Sc. degrees in electronic engineering from Sharif University of Technology, Tehran, Iran, in 2010 and 2012, respectively. He is currently pursuing the Ph.D. degree with the School of Communications and Networking, 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.

Hanan Al-Tous (SM' 2017) received the B.Sc. degree in Electrical Engineering in 1998 and the M.Sc. in Communication Engineering in 2000 from the University of Jordan, Amman, Jordan, and the Ph.D. degree in electrical engineering in 2014 from United Arab Emirates University, Al Ain, United Arab Emirates. She worked as a lecturer with the Electrical Engineering Department, Al-Ahliyya Amman University, Amman, Jordan from 2000-2010. She worked as a postdoc at the United Arab Emirates University, Al Ain, United Arab Emirates from 2014-2018. She is currently a research fellow at Aalto University, Finland. Her research interests include cooperative communications, energy harvested sensor networks, resource allocation for wireless communications, game theory, compressive sensing and machine learning.

Giuseppe Caire (S '92 – M '94 – SM '03 – F '05) was born in Torino in 1965. He received the B.Sc. in Electrical Engineering from Politecnico di Torino in 1990, the M.Sc. in Electrical Engineering from Princeton University in 1992, and the 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, and he is a recipient of the 2021 Leibniz Prize of the German National Science Foundation (DFG). 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 (SM'20) 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, from 1999 to 2010 he was with Nokia Research Center (NRC), Helsinki, Finland. 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. He is an editor of IEEE Transactions on Wireless Communications.