Optimal Bandwidth Allocation for Multicast-Cache-Aided on-Demand Streaming in Wireless Networks

Mohsen Amidzadeh*, Olav Tirkkonen* and Giuseppe Caire†
*Department of Communications and Networking, Aalto University, Espoo, Finland
†Communications and Information Theory Chair, TU Berlin, Germany
{mohsen.amidzade, olav.tirkkonen}@aalto.fi, giuseppe.caire@gmail.com

GLOBECOM 2022 - 2022 IEEE Global Communications Conference | 978-1-6654-3540-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/GLOBECOM48099.2022.10001076

***Abstract*—We consider a hybrid delivery scheme for streaming content, combining cache-enabled Orthogonal Multipoint Multicast (OMPMC) and on-demand Single-Point Unicast (SPUC) transmissions for heterogeneous networks. The OMPMC service transmits cached files through the whole network to interested users, and users not being satisfied by this service are assigned to the SPUC service to be individually served. The SPUC fetches the requested files from the core network and unicasts them to UEs using cellular beamforming transmissions. We optimize the delivery scheme to minimize the average resource consumption in the network. We formulate a constrained optimization problem over the cache placement and resource allocation of the OMPMC component, as well as the multi-user beamforming scheme of the SPUC component. We apply a path-following method to find the optimal traffic offloading solution. The solutions portray a contrast between the total amount of consumed resources and service outage probability. Simulation results show that the hybrid scheme provides a better tradeoff between the amount of network-wide consumed resources and the service outage probability, as compared to schemes from the literature.**

***Index Terms*—Hybrid content delivery, wireless caching, multipoint multicasting, single-point unicasting, resource consumption, zero-forcing beamforming, parametric optimization.**

I. Introduction

Wireless caching [1] is a candidate method to alleviate the unprecedented data congestion and traffic escalation issues in cellular networks. To determine a cache policy, the two phases of cache placement and cache delivery have to be defined [2], and both should be optimized to achieve a viable cache strategy.

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

For the content delivery, multipoint multicast (MPMC), single-point multicast (SPMC), and single-point unicast (SPUC) transmission schemes should be distinguished. MPMC utilizes multiple serving nodes to cooperatively broadcast files through the whole network, whereas the SPUC exploits on-demand transmissions to individually satisfy requesting User Equipments (UEs). For the SPMC scheme, each caching Base Station (BS) multicasts a file to several requesting UEs.

SPUC has been considered in [4], [6], [8] as an on-demand cache delivery scheme for HetNets. In [4], Zero Forcing Beamforming (ZFB) BSs and cache-equipped helper nodes were considered where the requesting UE is associated with the helper node with maximum received power. In [8], random resource allocation was utilized with SPUC. The fraction of bandwidth and the probabilistic cache placement were optimized to maximize offloading.

SPMC has been used in [9]–[11] as a content delivery method. In [9], [11], probabilistic caching is exploited for a two-tier HetNet where each BS multicasts k files in disjoint resources. In [10], coded caching is applied for a cellular network with multi-antenna BSs. The authors exploit joint SPUC and SPMC beamforming for content delivery, and the beamforming vectors are optimized to maximize the minimum rate of the UEs.

Digital Terrestrial TV Broadcasting systems [12] deliver content based on multipoint broadcast transmissions, while multipoint multicast (MPMC) delivery is used in the Long Term Evolution (LTE) system in the context of Multimedia Broadcast Multicast Services [13]. In a multi-cell transmission mode, all serving BSs stream the same file through the whole network in a Single-Frequency-Network (SFN) configuration. A SFN-configured MPMC was utilized in [14] for edge caching cellular networks. An orthogonal MPMC (OMPMC) was considered where resources are network-wide orthogonalized among cached files. The cache policy was then designed with respect to (w.r.t) file-specific resource allocation and probabilistic cache placement.

In this paper, we devise a hybrid content delivery scheme for cache-enabled multiantenna HetNets, combining OMPMC and SPUC. SPUC can serve individual users, but suffers from Co-Channel Interference (CCI). In contrast, OMPMC can serve a population of users interested in a given file with limited CCI [15]. However, this is obtained at the cost of the increase in the outage probability, as the multicast transmission can not individually satisfy UEs according to the resources need for

successful reception. Also, if only few users are interested in a file, using OMPMC would consume large amounts of network resources as compared to SPUC. This portrays a trade-off between OMPMC and SPUC delivery. We exploit this to develop a hybrid scheme based on OMPMC that streams the files using a network-wide file-specific resource allocation, and on-demand SPUC that unicasts the files using UE-specific resource-allocations. In contrast to [10], [16], we devise the hybrid scheme such that all UEs being dissatisfied by the multicast component can be properly served by the network. We optimize the network w.r.t. total amount of consumed resources subject to inevitable outages due to coping with the possibility of too large bandwidth requests.

More specifically, for the proposed hybrid SPUC/OMPMC delivery scheme, we find expressions for the outage probability and network resource consumption, using stochastic geometry. We then formulate the optimal traffic offloading problem as a joint optimization problem over cache placement, resource allocation and multiuser beamforming. To solve this problem, we represent it as a parametric optimization problem and leverage a path-following method to find the global optimal policy from the perspective of total resource consumption. We compare the proposed hybrid scheme to cache delivery policies from the literature, in different settings.

The remainder of this paper is organized as follows. In Section II, the system model is introduced. In Section III, the total resource consumption of the considered hybrid scheme is computed. The optimization problem is formulated in Section IV, while simulation results are presented and discussed in Section V. Finally, Section VI concludes the paper.

Notations: In this paper, we use lower-case a for scalars, bold-face lower-case $\mathbf{a}$ for vectors and bold-face uppercase $\mathbf{A}$ for matrices. Further, $\{a_n\}_1^N$ collects the components of vector $\boldsymbol{a}$ from $n=1$ to $n=N$. We use $\mathbf{1}$ and $\mathbf{0}$ to denote the vector with all elements equal to one and zero, respectively. We use $\dot{\boldsymbol{a}}(\theta)$ to represent the derivative of $\boldsymbol{a}(\theta)$ with respect to θ.

II. System Model

We consider a content library containing N different files. The fraction of active UEs requesting file n is f_n. We assume that file popularities $\{f_n\}_1^N$ are known to the network. We use the Zipf distribution (see [17]) to model popularity; $f_n = n^{-\tau}/\sum_{m=1}^{N} m^{-\tau}$, with τ being the skewness of the Zipf distribution. We consider a streaming service and for simplicity, we assume the same streaming rate R for all files which leads to the identical service quality. The study of dynamic rate/quality is left for future work.

We consider a two-tier HetNet with BSs and Helper Nodes (HNs). The network applies a hybrid delivery scheme, based on OMPMC and SPUC components to serve UEs. The HNs constitute the multicast component and are equipped with limited-capacity caches which proactively cache files based on a probabilistic approach [18]. The HNs apply OMPMC to broadcast the cached files through the whole network in file-specific disjoint resources. In the OMPMC layer, there may be co-channel interference arising from far-away HNs. However, with conventional Orthogonal Frequency Division Multiplexing (OFDM) parameterization, such co-channel interference would be negligible [15]. The BSs constitute the unicast component. They are connected to a high-capacity backhaul link which can fetch files from the core-network, and serve these to users using on-demand unicast. The BSs are equipped with L antennas and use multiuser ZFB to serve u users simultaneously in a frequency resource.

The network dedicates network-wide disjoint resources for the OMPMC and SPUC delivery components, denoted by W^{MC} and W^{UC}, respectively. Each BS uses the full bandwidth W^{UC}, i.e., frequency reuse 1 is used in the unicast layer. When served by the unicast layer, a UE is connected to the nearest BS. To support the streaming rate R, the BS allocates a sufficient amount of resources to a UE, if it is not greater than a service threshold.

We assume that the locations of UEs, BSs and HNs are based on three independent Poisson Point Processes (PPPs), i.e., Φ_u and Φ_b and Φ_h, respectively, with intensities λ_u, λ_b and λ_h.

A. Cache Placement and Content Fetching

The HNs are equipped with caches of limited memory capacity so that they can store at most M files. They proactively and independently cache files based on a probabilistic approach [3] modeled by a common probability distribution. As such, cache weights $\{p_n\}_1^N$ with $0 \leq p_n \leq 1$ and $\sum_{n=1}^{N} p_n \leq M$ are introduced, and an approach is devised such that each HN can store exactly M files with file n being cached at the HN with probability equal to p_n. Note that the files are cached at HNs based on another PPP with intensity $\{\lambda_u p_n\}_1^N$, according to the thinning property of Φ_u [19].

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

B. Content Delivery

The network operates in a time-slotted fashion. Within each time-slot, requests arrive at arbitrary times from a spatial realization of UEs based on Φ_u. The network applies a hybrid delivery scheme based on OMPMC and SPUC transmissions to satisfy these requests. Any UE not being satisfied by the multicast component is assigned to the unicast component to be properly served.

The multicast component streams cached files through whole network by cooperation of all HNs. It uses file-specific disjoint resources for broadcasting different files bandwidth w_n^{MC} allocated for file n. To reduce the latency of streaming in the OMPMC mode, we apply Harmonic Broadcasting (HB) characterized by harmonic number H_w [20], though the analysis is not restricted to any particular streaming scheme. When the consumed bandwidth is increased by a factor of H_w, the average latency of a user requesting the file within the time slot is reduced by a factor of $1/D$, where $H_w = \sum_{i=1}^{D} 1/i$.

We assume that the average transmission power of all HNs is the same, denoted by $p_{\rm tx}$. With equal power allocation across frequency, the transmitting Signal-to-Noise-Ratio (SNR) of all files in OMPMC are the same, $\gamma_{\rm tx}^{\rm MC} = \frac{p_{\rm tx}}{W^{\rm MC} N_0}$, where N_0 is the noise spectral density.

Any UE not being satisfied by the multicast component, requests the file from the unicast component. As such, the BSs in the unicast layer, constitute a Poisson-Voronoi tessellation with different cell sizes. The nearest BS to the UE fetches the file from the core network and unicasts it towards the UE. If the Signal-to-Interference-plus-Noise Ratio (SINR) of a user is above a threshold, the responsible BS responds to the UE by allocating the resources it needs to successfully decode its files. In each Voronoi cell, there may be multiple UEs requesting files.

We assume that the average transmission power of all BSs is $p_{\rm tx}$, and they are equipped with L antennas and serve u UEs at the same frequency bandwidth using a ZFB transmission.

III. Resource Consumption Analysis

We consider separately the multicast and the unicast components, and then provide an expression for the total resource consumption of the hybrid system.

A. Resource Consumption of Multicast Component

The multicast component applies network-wide resource allocation with OFDM-based transmission to broadcast the files through the network. The Signal-to-Noise-Ratio (SNR) of UE k requesting file n is expressed as [14]:

$$\gamma_{k,n}^{\rm MC} = \gamma_{\rm tx}^{\rm MC} \sum_{j \in \Phi_{p,n}} |h_{j,k}|^2 \|\boldsymbol{x}_k - \boldsymbol{r}_j\|^{-e}, \tag{1}$$

where $\Phi_{p,n}$ stands for the set of HNs caching file n, $h_{j,k}$ is the channel coefficient between HN j and UE k, $\boldsymbol{x}_k$ and $\boldsymbol{r}_j$ are the locations of UE k and HN j, respectively, and e is the path-loss exponent. We use a standard distance-dependent to model the path-loss, and assume an Rayleigh distribution for the channel coefficient, i.e., $|h_{j,k}|^2 \sim \exp(1)$.

The maximum achievable rate for a transmission is obtained from the corresponding Additive-White-Gaussian-Noise channel capacity. If the maximum rate experienced by a UE is less than the streaming rate R, the UE is in outage. The outage probability $\mathcal{O}_{n,k}^{\rm MC}$ for UE k requesting file n then is:

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

Defining a spectral efficiency threshold $\alpha_n = R/w_n^{\rm MC}$, the total resource usage of the multicast component is $W^{\rm MC} = \sum_{n=1}^{N} w_n^{\rm MC} = \sum_{n=1}^{N} \frac{R}{\alpha_n}$. However, as we apply harmonic broadcast [20], the effective resource usage will be multiplied by H_w, i.e.,

$$W_{\rm eff}^{\rm MC}(\boldsymbol{\alpha}) = H_w \sum_{n=1}^{N} \frac{R}{\alpha_n}.$$

Based on the Slivnyak-Mecke theorem [19], the performance can be computed for a typical UE located at the origin. Hence, we can set $\mathcal{O}_{n,0}^{\rm MC} = \mathcal{O}_n^{\rm MC}$. Accordingly, the outage probability of file n, served by the multicast component, is [14]:

$$\mathcal{O}_n^{\rm MC}(p_n, \{\alpha_l\}_l) = \frac{2}{\pi} \int_0^\infty \left\{ \frac{1}{w} \cos\left(\frac{\pi^2 \lambda_h p_n}{e \sin(\pi/e)} \left(\frac{w}{g_{\boldsymbol{\alpha}}} \right)^{2/e} \right) \times \exp\left(\frac{-\pi^2 \lambda_h p_n}{e \sin(\pi/e)} \left(\frac{w}{g_{\boldsymbol{\alpha}}} \right)^{2/e} \right) \sin\left(\frac{w}{\gamma_R} \right) \right\} dw, \tag{2}$$

where $g_{\boldsymbol{\alpha}} = \left(2^{\alpha_n} - 1\right) \sum_{l=1}^{N} \alpha_l^{-1}$ and $\gamma_R = \frac{p_{\rm tx}}{N_0} \frac{1}{R H_w}$.

B. Resource Consumption of Unicast Component

In each Voronoi cell of the unicast layer, zero-forcing beamforming is utilized to serve u UEs in the same bandwidth. Consequently, for UE k served by the unicast component, the SINR is:

$$\gamma_k^{\rm UC} = \frac{g_k \|\boldsymbol{x}_k - \boldsymbol{r}_0\|^{-e}}{1/\gamma_{\rm tx}^{\rm UC} + \underbrace{\sum_{j \in \Phi_b / \boldsymbol{r}_0} g_j^k \|\boldsymbol{x}_k - \boldsymbol{r}_j\|^{-e}}_{I_k}} \tag{3}$$

where $\gamma_{\rm tx}^{\rm UC} = p_{\rm tx}/(W^{\rm UC} N_0)$, g_k is the effective channel gain between the nearest BS and UE k, constructed from the channel vector and the beamforming vector, and g_j^k is the effective channel gain from BS j to UE k. Further, $\boldsymbol{r}_0$ and $\boldsymbol{r}_j$ are the locations of the nearest BS and BS j. Consequently, we have $g_k \sim \Gamma(L - u + 1, 1)$ and $g_j^k \sim \Gamma(u, 1)$ [21], with $\Gamma(a, b)$ being the gamma distribution with shape a and scale b.

In each cell, the BS allocates a sufficient amount of resources to the served UEs. However, if the SINR of a user is very small, near infinite bandwidth would be needed to serve the user. To cope with this, we apply a thresholding policy, the bandwidth allocated to serve UE k is:

$$w_k^{\rm UC} = \begin{cases} R/\log_2(1 + \gamma_k^{\rm UC}), & \gamma_k^{\rm UC} \ge \gamma_{\rm th} \\ 0, & \text{otherwise} \end{cases}, \tag{4}$$

where $\gamma_{\rm th}$ is the SINR threshold. Hence, no resource is allocated if the UE is in a poor condition. Considering that each BS applies multiuser ZFB towards u users, the resources consumed in each cell is approximately given by the total amount of resources allocated to UEs in that cell divided by u. This becomes precise when $\lambda_u/\lambda_b \to \infty$, and is a sufficient approximation when $\lambda_u \gg \lambda_b$. We assume that all BSs are active during unicast service, so the average resource consumption by this component is determined by the average consumption of a typical Voronoi cell $\mathcal{V}_0$:

$$W^{\rm UC}(\cdot) = \mathbb{E}\left\{ \sum_{k \in \mathcal{V}_0} \frac{w_k^{\rm UC}}{u} \right\} \tag{5}$$

The expectation is w.r.t. Poisson processes of UEs (Φ_u) and BSs (Φ_b), as well as effective channel gains based on (3).

For the resource consumption in the unicast layer we have:

Theorem 1. *Consider an interference-limited frequency reuse 1 cellular network with BSs and UEs coming from PPPs with intensities λ_b and λ_u^{eff}, respectively, where UEs are served by their nearest BS. BSs use ZFB with L antennas towards u simultaneous users, and allocate bandwidth to users using service threshold γ_{th} as in (4). The average amount of resources needed is then:*

$$W^{\text{UC}}(u,\lambda_u^{\text{eff}}) = \frac{\lambda_u^{\text{eff}}}{2u\lambda_b} \sum_{l=1}^{L-u+1} f_{l,u} \int_0^{w_{\text{th}}} w \frac{d}{dw} C_{l,u}(w) dw, \quad (6)$$

where

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

$w_{\text{th}} = R/\log_2(1+\gamma_{\text{th}})$, $\eta(w) = 2^{R/w} - 1$, $\xi = (L-u+1)!^{\frac{-1}{L-u+1}}$, $f_{l,u} = (-1)^{l+1}\binom{L-u+1}{l}$, *and* ${}_2F_1(\cdot)$ *is the hypergeometric function. The outage probability of a UE is:*

$$\mathcal{O}^{\text{UC}}(u,\gamma_{\text{th}}) = 1 - \sum_{l=1}^{L-u+1} \frac{f_{l,u}}{{}_2F_1(-\frac{2}{e}, u, 1-\frac{2}{e}, -\xi\, l\, \gamma_{\text{th}})}.$$

Proof. It is available in a preprint version of the paper [22]. □

Notice that $W^{\text{UC}}(\cdot)$ depends on $\lambda_u^{\text{eff}}/\lambda_b$, not on both λ_b and λ_u^{eff} separately.

C. Resource Consumption of the Hybrid Scheme

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

$$\frac{dW^{\text{UC}}(u,\lambda_u^{\text{eff}})}{d\gamma_{\text{th}}} = -\frac{w_{\text{th}}\,\lambda_u^{\text{eff}}}{2u\,\lambda_b} \frac{d\mathcal{O}^{\text{UC}}(u,\gamma_{\text{th}})}{d\gamma_{\text{th}}} \quad (7)$$

Equation (7) directly provides a trade-off between the outage probability, and resource consumption of the unicast scheme. Considering that $\frac{d\mathcal{O}^{\text{UC}}(u,\gamma_{\text{th}})}{d\gamma_{\text{th}}} > 0$, increase in w_{th} makes W^{UC} grows monotonically.

Not all UEs request from the unicast component; only UEs being dissatisfied by the multicast component do. Based on the thinning property of PPP, λ_u^{eff} in (6) depends on the overall outage probability of the multicast component: $\mathcal{O}^{\text{MC}}(\boldsymbol{p},\boldsymbol{\alpha}) = \sum_{n=1}^{N} f_n \mathcal{O}_n^{\text{MC}}(p_n, \{\alpha_l\}_l)$, considering that f_n is the popularity of file n. Therefore, the average total resource consumption of the whole network is expressed as:

$$W_{\text{tot}} = \underbrace{H_w R \sum_{n=1}^{N} \alpha_n^{-1}}_{\text{multicast component}} + \underbrace{W^{\text{UC}}(u,\lambda_u)\mathcal{O}^{\text{MC}}(\boldsymbol{p},\boldsymbol{\alpha})}_{\text{unicast component}}. \quad (8)$$

Accordingly, we can obtain the service outage probability, defined as the probability that a typical UE being served by the hybrid scheme is in outage. It depends on the outage of the SPUC and OMPMC components as:

$$\mathcal{O}_{\text{tot}} = \mathcal{O}^{\text{UC}}(u,\gamma_{\text{th}})\,\mathcal{O}^{\text{MC}}(\boldsymbol{p},\boldsymbol{\alpha}).$$

IV. Optimal Traffic Offloading Policy

Aiming to minimize the total resources consumed by overall on-demand streaming service, we minimize the total resource consumption of the hybrid scheme with respect to the caching policy (probabilities $\boldsymbol{p}$), the inverse of the spectral efficiency thresholds of the multicast component $\{\beta_n := \frac{1}{\alpha_n}\}_1^N$, and the multiuser MIMO spatial multiplexing u of the cellular base stations. This yields the problem:

$$P_1: \min_{\boldsymbol{p},\boldsymbol{\beta},u} H_w R \sum_{n=1}^{N} \beta_n + W^{\text{UC}}(u,\lambda_u) \sum_{n=1}^{N} f_n \mathcal{O}_n^{\text{MC}}\left(p_n, \left\{\frac{1}{\beta_l}\right\}_l\right)$$
$$\text{s.t.} \begin{cases} \beta_n \geq 0, & 0 \leq p_n \leq 1, \qquad n \in S_N, \\ \sum_{n=1}^{N} p_n \leq M, & u \in \{1,\dots,L\}, \end{cases}$$

where $S_N = \{1,\dots,N\}$.

Note that P_1 is a non-convex mixed-integer optimization problem. The optimization parameters are the cache weights and resource allocations of the multicast component, i.e., $\{(p_n,\beta_n)\}_1^N$, and the multiuser beamforming parameter u of the unicast component .

To optimize P_1, w.r.t. u, we use a simple line search with $u \in \{1,\dots,L\}$. However, to find the optimum value for $\{(p_n,\beta_n)\}_1^N$, we use a path-following method [23]. As such, we formulate a parametric optimization problem exactly as P_1 but with θ-parameterized popularity $b_n(\theta)$ replacing f_n. We denote the corresponding parametric optimization problem by $P_1(\theta)$. We parameterize $b_n(\theta)$ such that $\lim_{\theta\to 0} b_n(\theta) = \frac{1}{N}$ and $\lim_{\theta\to\tau} b_n(\theta) = f_n$, for $n \in S_N$. As such, the solution of $P_1(\theta)$ can be found by solving ODE (9), written at bottom of next page, where $i,j \in S_N$, δ_{ij} is the Kronecker delta function, $[a_{ij}]_{i,j}$ constitutes a matrix whose i-th row and j-th column is a_{ij} and $[a_i]_i$ is a column vector with i-th component being a_i. We solve ODE (9) for $\theta \in [0,\tau]$ by the sequential method elaborated in [15]. It gives the solution of the sough optimization problem $P_1(\tau)$ of interest, corresponding to the target popularities $\{f_n\}_1^N$. However, an initial point, at $\theta = 0$, is needed for solving this ODE. To obtain it, we consider the optimization problem $P_1(0)$ corresponding to the popularity $b_n(0)$ and find its optimal solution. For the parameterization, we set $b_n(\theta) = n^{-\theta}/\sum_{m=1}^{N} m^{-\theta}$ for $n \in S_N$ and $\theta \in [0,\tau]$.

V. Simulation Results and Discussion

We compare the optimal cache delivery solution of the hybrid scheme to conventional multi-antenna SPUC [4], [21], [24] and OMPMC schemes [14] from the literature.

We consider the following scenario. The number of files is $N = 100$, the popularity skewness is $\theta = 0.6$, the cache capacity of BSs is $M = 10$, the streaming service rate $R = 1$ Mbps, and having $L = 8$ antennas at the BSs and the service threshold $\gamma_{\text{th}} = 0.1$, unless it is specified. We set the Harmonic number $H_w = 6$, which approximately reduces the streaming latency with a factor of 226 [20]. The BSs and HNs are deployed based on two independent homogeneous PPP with

intensity $\lambda_b = 200$ and $\lambda_h \in [20, 200]$, respectively, and UEs are located according to another homogeneous PPP with intensity $\lambda_u = 2 \times 10^5$. We apply an Urban NLOS scenario from 3GPP [25] with carrier frequency 2 GHz, HN transmission power 23 dBm, and path-loss exponent $e = 3.76$. The network of BSs is assumed interference-limited. The antenna gain at the UE and HN are 0 dBi and 8 dBi, respectively, the noise-figure of UE is 9 dB, the noise spectrum density is -174 dBm. Note that since the reference distance is 1 km, the UE and BS intensities are in the units of points/km^2.

Figure 1 shows the normalized resource usage for the hybrid scheme, and the SPUC and OMPMC components as a function of HN intensity λ_h. As λ_h increases, the total resource consumption decreases–the OMPMC component consumes more resources, as more files are offloaded to be served in OMPMC.

Figure 2 illustrates the outage probability as a function of HN intensity λ_h. Note that the outage probability of the SPUC component is insensitive to λ_h. As λ_h increases, the total outage probability of the hybrid scheme decreases due to the increase in the performance of the OMPMC component.

To investigate how the proposed hybrid scheme optimizes the total resource usage of the network, we compare it with OMPMC and SPUC delivery schemes. Figure 3 portrays the tradeoff between total resource usage and outage probability of the different schemes for $\lambda_b = 200$, and a sparse MPMC component with $\lambda_h = 50$. To generate the curves of SPUC and hybrid schemes, we change the service threshold in the range $\gamma_{\mathrm{th}} \in [0.03, 1]$. For the OMPMC scheme, we target a value for the outage probability and determine the corresponding consumed resources. The OMPMC scheme performs worst as compared to the SPUC and the hybrid schemes. The reason for this is that the considered Zipf distribution with $\tau = 0.6$ has a fat tail—there is a significant number of request for non-cached files. Despite this, the hybrid scheme is able to considerably outperform the SPUC scheme, offloading the most popular files to the OMPMC component.

The same tradeoff for a cellular network, where $\lambda_b = \lambda_h = 200$, is depicted in Figure 4. In this case, the BS and caching HNs can be considered to be the same. Now, OMPMC generically outperforms SPUC, but the hybrid scheme still provides the best service with a wide margin.

VI. Conclusion

In this paper, we considered hybrid content delivery combining orthogonal multipoint multicast (OMPMC) and single-point unicast (SPUC) transmission schemes. A traffic offloading policy jointly optimizing cache placement and radio resource allocation of OMPMC and multiuser beamforming of SPUC was formulated, based on the derived expression of network-wide consumed resource. The amount of consumed resources depends on the ratio of user intensity to SPUC BS intensity, and not on both separately. Further, a functional relationship between the total resource consumption and the outage probability of unicast service was derived. To find the optimal policy, a parametric optimization problem was formulated and solved using a path-following method. We compared the performance of the hybrid scheme with using only SPUC or OMPMC. Simulation results clearly portray the tradeoff between total resource consumption and service outage probability, and showed that despite a fat tail of popularity distribution, which renders OMPMC ineffective as compared to SPUC, the hybrid scheme outperforms SPUC with a wide margin, in the simulated scenarios up to 50% of the resources can be saved by using the hybrid scheme instead of SPUC. The proposed hybrid delivery scheme is a promising candidate for optimizing the spectral efficiency and total resource consumption in heterogeneous and cellular networks.

Acknowledgment

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

References

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

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

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

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

[5] 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.

[6] 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.

$$\begin{pmatrix} \dot{\boldsymbol{\beta}}(\theta) \\ \dot{\boldsymbol{p}}(\theta) \\ v(\theta) \end{pmatrix} = - \begin{pmatrix} \left[b_i(\theta) \frac{d^2 \mathcal{O}_i^{\mathrm{MC}}}{dp_i d\beta_j} \right]_{i,j} & \left[b_i(\theta) \frac{d^2 \mathcal{O}_i^{\mathrm{MC}}}{dp_i^2} \delta_{ij} \right]_{i,j} & \mathbf{1} \\ \left[\sum_{n=1}^{N} b_n(\theta) \frac{d^2 \mathcal{O}_n^{\mathrm{MC}}}{d\beta_i d\beta_j} \right]_{i,j} & \left[b_j(\theta) \frac{d^2 \mathcal{O}_j^{\mathrm{MC}}}{dp_j d\beta_i} \right]_{i,j} & \mathbf{0} \\ \mathbf{0}^\top & \mathbf{1}^\top & 0 \end{pmatrix}^{-1} \begin{pmatrix} \left[\dot{b}_i(\theta) \frac{d\mathcal{O}_i^{\mathrm{MC}}}{dp_i} \right]_i \\ \left[\sum_{n=1}^{N} \dot{b}_n(\theta) \frac{d\mathcal{O}_n^{\mathrm{MC}}}{d\beta_i} \right]_i \\ 0 \end{pmatrix} \tag{9}$$

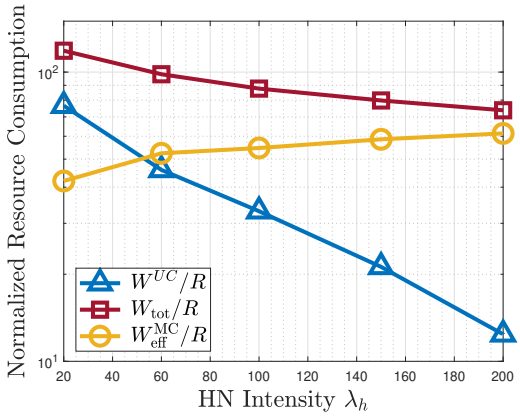

Fig. 1: The normalized resource consumption as a function of HN intensity λ_h for $\lambda_b = 200$.

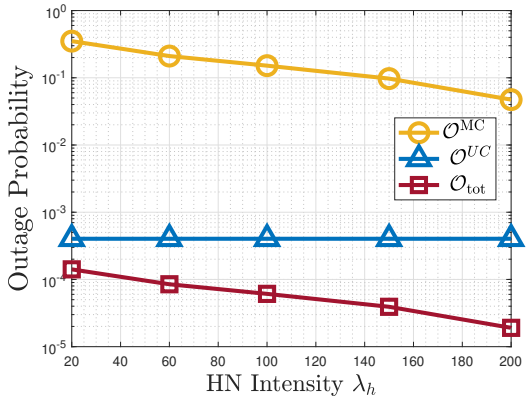

Fig. 2: The outage probability as a function of HN intensity λ_h for $\lambda_b = 200$.

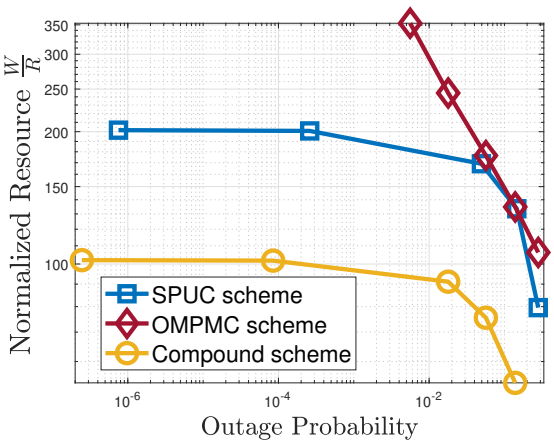

Fig. 3: Normalized resource consumption versus outage probability for $\lambda_b = 200$, $\lambda_h = 50$.

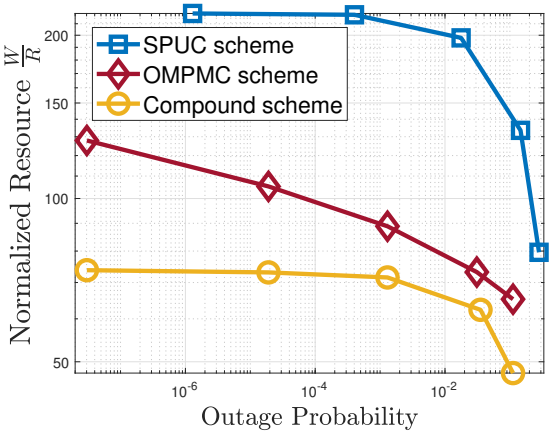

Fig. 4: The normalized resource versus outage probability for $\lambda_b = \lambda_h = 200$.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

[16] 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 *Proc. IEEE Wireless Commun. and Netw. Conf. (WCNC)*, Dec. 2022, pp. 1–6.

[17] 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.

[18] 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.

[19] 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.

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

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

[22] M. Amidzadeh, G. Caire, and O. Tirkkonen, "Optimal bandwidth allocation for multicast-cache-aided on-demand streaming in wireless networks," preprint ArXiv 2205.03189, 2022.

[23] 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.

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

[25] 3GPP, "Universal mobile telecommunications system (UMTS); radio frequency RF system scenarios," 3rd Generation Partnership Project (3GPP), Technical Report (TR), April 2017, version 14.0.0.