Optimal Resource Allocation for NOMA-Enabled Cache Replacement and Content Delivery

Lei Lei[1], Thang X. Vu[1], Lin Xiang[1], Xingjun Zhang[2], Symeon Chatzinotas[1], and Björn Ottersten[1]

[1]Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg
[2]Department of Computer Science and Technology, Xian Jiaotong University, Xi'an, China
Emails: {*lei.lei;thang.vu; lin.xiang; symeon.chatzinotas; bjorn.ottersten@uni.lu, xjzhang@mail.xjtu.edu.cn*}

***Abstract*—In a content-delivery network, files' popularity and users' requests change fast. Conventional caching schemes, e.g., caching (re)placement once per day during the off-peak hours, may not capture the up-to-date popularity. In this case, the contents in caches have to be regularly updated to prevent information becoming outdated, and at the same time users' requested files must be delivered. These two tasks are challenging in practical heavy-traffic and multi-user scenarios when the network resources are limited. In this paper, we apply non-orthogonal multiple access (NOMA) to facilitate concurrent caching replacement and content delivery in downlink transmission. We formulate a resource allocation problem to investigate how to efficiently push proactive files to the cache at the small base station and deliver the requested files to users. The resource-allocation problem is formulated as a mixed-integer exponential conic optimization problem. To enable a computationally-efficient optimal solution with finite convergence, we develop an iterative algorithm based on polyhedral outer approximation, where a polyhedral relaxation subproblem and a convex subproblem are constructed and iteratively solved to tighten the lower and upper bounds for the optimum, respectively. The numerical results demonstrate significant performance gains of the NOMA-enabled data transmission scheme in power and resource savings compared to the baseline scheme.**

***Index Terms*—NOMA, caching, resource allocation, mixed-integer convex programming.**

I. Introduction

In cache-enabled networks, contents' generation, popularity, and requests are highly dynamic. Firstly, a large amount of new contents, e.g., popular news, video, are generated over time. Secondly, the average lifetime of a popular file may change fast. For example, the popularity of an up-to-date file could diminish within seconds, minutes, or hours [1]. Thirdly, the users' requests and preference may also change frequently. Content updating in a caching system is typically carried out infrequently, e.g., cache (re)placement during the mid-night when the network resources are idle [1]. This solution is applicable if the above three aspects vary slowly. It may not be applicable for practical peak-hour scenarios. A potential issue is that if the cached content cannot be updated timely, when the requests for a new content with high popularity arrive, the system has to wait until the next cache replacement to deliver it to the cache. During this long time interval, the cache hit probability could progressively decrease. More users may have to request this new content from the macro base station (MBS) remotely. As a result, more resources, e.g., power/energy, time, and channels, need to be consumed than serving users directly from the local cache. Therefore, in-time cache replacement is important for peak-hour scenarios. Two realistic problems, strong interference and limited available resources, have to be tackled in simultaneous cache replacement and content delivery. Non-orthogonal multiple access (NOMA) can break the orthogonality in resource allocation and allow successive interference cancellation (SIC) to mitigate co-channel interference. Thus it provides an opportunity to enable concurrent cache update and content delivery, and an alternative trade-off between the above two issues.

For efficient caching (re)placement, in [2], an enhanced transmission scheme was proposed to enable energy-efficient content placement. The authors in [3] developed an online algorithm with light overhead for fast cache replacement. On the other hand for efficient content delivery, the authors in [4] investigated minimum-latency optimization in caching content delivery. In [5], the authors proposed optimal and suboptimal solutions for minimum-energy content delivery in multi-carrier multi-cell networks. Recently NOMA has been increasingly considered in caching systems. In [6], a NOMA assisted caching scheme was proposed to improve the cache hit probability and reduce the delivery outage probability. The authors in [7] investigated simultaneous caching and content multicasting, and the expressions of outage probability were derived. In [8], the authors studied NOMA-based content delivery with considering service deadlines. In [9], a cache-aided NOMA scheme was proposed to exploit cached data for interference cancellation. As an emerging application area of NOMA and caching, the optimal resource allocation for joint cache update and content delivery is not adequately studied. To bridge this

gap, in this paper we provide optimal solutions for how to use the limited system resources to complete cache replacement and content delivery efficiently by consuming less power. The problem is formulated as a mixed-integer exponential conic optimization problem. A polyhedral outer approximation algorithm is developed to enable global optimum. By comparing the optimal solutions between a NOMA-based scheme and an orthogonal multiple access (OMA) based scheme in short-term and long-term performance, the numerical results show the performance improvements of the developed solution.

II. System Model

A. Cache Update and Content Delivery

We consider downlink transmission in a two-tier heterogeneous cellular system, where an MBS and a small base station (SBS) are deployed to serve up to K user equipments (UEs). The set of UEs is denoted as $\mathcal{K}$. The SBS operates in the half-duplex mode (HD-SBS), and is equipped with a storage-limited cache. Let $\mathcal{F}_c$ be the set of files stored in the cache, and $\mathcal{F}_u$ be the set of files that are currently requested by the UEs. The SBS transmits a requested file f to its associated UE directly if f is available in $\mathcal{F}_c$. Otherwise, f has to be transmitted via the MBS-to-UE link. The MBS is able to retrieve any file from the core network when the file is not cached at the SBS. We denote the sets of files to be transmitted from the SBS and the MBS to UEs as $\mathcal{F}_s = \mathcal{F}_c \cap \mathcal{F}_u$ and $\mathcal{F}_m = \mathcal{F}_u \backslash \mathcal{F}_s$, respectively. Correspondingly, the sets of the receivers served by the SBS and the MBS are denoted as $\mathcal{K}_s$ and $\mathcal{K}_m$, respectively. In data transmission, the entire frequency band is divided to N subchannels, each of bandwidth B Hz. The time domain is divided into scheduling frames. Each frame consists of T time slots. As shown in Fig. 1, we define one subchannel with one time slot as a time-frequency resource unit (RU) which is the minimum unit to be allocated. The channel fading coefficients are fixed during a scheduling frame.

In the considered scenarios, all the UEs' requests $\mathcal{F}_u$ must be satisfied, and the cached content needs to be updated periodically in order to capture the up-to-date popularity. Thus the files to be transmitted can be divided into two sets, $\mathcal{F}_u$ and $\mathcal{F}_{ms}$. The former, $\mathcal{F}_u = \mathcal{F}_s \cup \mathcal{F}_m$, is currently requested by the UEs, and to be transmitted via the MBS-to-UE or SBS-to-UE links. $\mathcal{F}_{ms}$ contains the proactive-caching files which may not be demanded by UEs at the moment but could be requested with high probability in the near future. The cache hit probability could be progressively improved by periodically replacing $\mathcal{F}_c$ by $\mathcal{F}_{ms}$. In practice, some proactive caching techniques, e.g., machine learning based popularity prediction [10], can be used to obtain $\mathcal{F}_{ms}$. In the paper, how to precisely identify the files in $\mathcal{F}_{ms}$ is out of the scope of this study. We assume that both $\mathcal{F}_u$ and $\mathcal{F}_{ms}$, as the inputs, are known prior to the resource optimization. We focus on how to deliver all the files of $\mathcal{F}_u$ and $\mathcal{F}_{ms}$ in an efficient way. To push the files of $\mathcal{F}_{ms}$ to a fully-loaded cache, deciding which existing files in $\mathcal{F}_c$ should be moved out from the cache can be based on some well-known cache replacement policies, e.g., least recently used (LRU) [3].

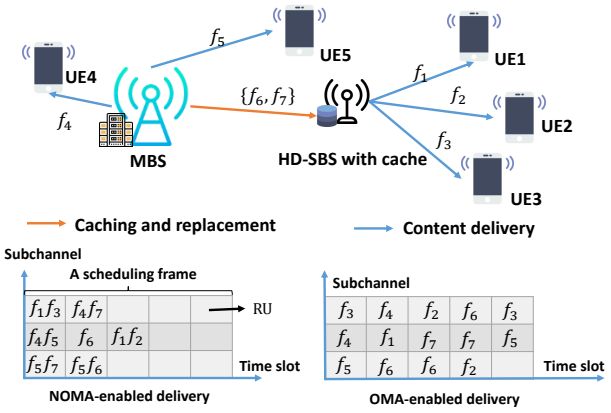

Figure 1. An illustrative example of resource allocation in cache replacement and content delivery. Files $f_1, \ldots, f_5$ are currently requested by UEs, where $\mathcal{F}_c = \{f_1, f_2, f_3\}$. The cached content needs to be updated by $\mathcal{F}_{ms} = \{f_6, f_7\}$. Applying NOMA-enabled delivery, only fewer RUs and times slots are required.

B. NOMA-Enabled Data Transmission

To mitigate the co-channel interference and improve the resource utilization, we exploit NOMA in both data transmission from the SBS to UEs, and from the MBS to UEs and the SBS. For example in Fig. 1, NOMA is applied within each group $\mathcal{F}_s = \{f_1, f_2, f_3\}$ and $\mathcal{F}_m \cup \mathcal{F}_{ms} = \{f_4, f_5, f_6, f_7\}$, where UEs within the same set can share the same RU. Superposition coding is applied at the transmitter (MBS or SBS) along with the application of SIC at the receiver side (UE or SBS). Two remarks are in order. Firstly, due to the half-duplex operation at the SBS, concurrent transmission among the links MBS-to-SBS and SBS-to-UEs are prohibited. Secondly, paring UE $k \in \mathcal{K}_m$ with UE $k' \in \mathcal{K}_s$ on the same RU is suboptimal due to the strong mutual interference and without interference cancellation. Hence it will be excluded in the optimum.

According to the NOMA basis [6], the optimal decoding order on a channel is predefined as the descending order of channel gains. For the SBS-associated UEs $k \in \mathcal{K}_s$ on channel n, we sort channel coefficients $|h^s_{1n}|^2 \ldots, |h^s_{kn}|^2 \ldots |h^s_{|\mathcal{K}_s|n}|^2$ by the descending order, where h^s_{kn} is the channel coefficient from the SBS to the kth UE in set $\mathcal{K}_s$ on channel n. For convenience, we assume $|h^s_{1n}|^2 \geq |h^s_{2n}|^2 \geq \cdots \geq |h^s_{kn}|^2 \geq \cdots \geq |h^s_{|\mathcal{K}_s|n}|^2$. The signal-to-interference-plus-noise ratio (SINR) for UE $k \in \mathcal{K}_s$ on channel n is given as,

$$\text{SINR}^s_{knt} = \frac{p^s_{knt}|h^s_{kn}|^2}{\sum_{k'=1}^{k-1} p^s_{k'nt}|h^s_{kn}|^2 + \sigma^2}, \quad k \in \{1, \ldots, |\mathcal{K}_s|\} \tag{1}$$

where p^s_{knt} is the transmit power for UE $k \in \mathcal{K}_s$ on channel n and time slot t, and σ^2 is the noise power. The kth UE in $\mathcal{K}_s$ before decoding its desired signal, firstly decodes and subtracts the interfering signals from the UEs $k+1$ to $|\mathcal{K}_s|$. On the other hand the signals intended for the first UE to the $k-1$th UE are treated as noise. The achievable data rates of UE $k \in \mathcal{K}_s$ on channel n and slot t is $r^s_{knt} = B\log_2\left(1+\mathrm{SINR}^s_{knt}\right)$.

Now consider NOMA-based delivery for the files in set $\mathcal{F}_m \cup \mathcal{F}_{ms}$. Without loss of generality, we assume $|h^m_{0n}|^2 \geq |h^m_{1n}|^2 \geq \cdots \geq |h^m_{\bar{k}n}|^2 \geq \cdots \geq |h^m_{|\mathcal{K}_m|n}|^2$. In $\mathcal{K}_m$, we treat the SBS as a special UE, $\bar{k}=0$, to receive the proactive caching files, and h^m_{0n} is the channel coefficient of the MBS-to-SBS link on channel n, and $h^m_{\bar{k}n}$ is the channel coefficient between the MBS and the $\bar{k}$th UE in set $\mathcal{K}_m$ on channel n,

$$\mathrm{SINR}^m_{\bar{k}nt} = \frac{p^m_{\bar{k}nt}|h^m_{\bar{k}n}|^2}{\sum_{\bar{k}'=0}^{\bar{k}-1} p^m_{\bar{k}'nt}|h^m_{\bar{k}n}|^2+\sigma^2},\ \bar{k}\in\{0,\ldots,|\mathcal{K}_m|\} \tag{2}$$

where $p^m_{\bar{k}nt}$ is the MBS's transmit power allocated for the $\bar{k}$th UE on channel n and time slot t. Note that the summation of interference $\sum_{\bar{k}'=0}^{\bar{k}-1} p^m_{\bar{k}'nt}|h^m_{\bar{k}n}|^2 = 0$ for $\bar{k}=0$ in the current decoding order. The achievable data rates of the $\bar{k}$th UE in $\mathcal{K}_m = \{0,\ldots,|\mathcal{K}_m|\}$ is $r^m_{\bar{k}nt} = B\log_2\left(1+\mathrm{SINR}^m_{\bar{k}nt}\right)$.

III. Problem Formulation

In this section, we address a joint optimization problem for cache replacement and content delivery, aiming at delivering all the required data in a resource-constrained and multi-UE scenario. The optimization task is to allocate time, frequency, and power resources to satisfy all the UEs' requests and deliver all the proactive caching files in an efficient manner. We formulate the resource allocation problem in P0. We introduce three sets of variables, i.e., continuous variables $p^s_{knt}, k\in\mathcal{K}_s$ and $p^m_{\bar{k}nt}, k\in\mathcal{K}_m$ to represent power allocation, binary variables $x^s_{knt}, k\in\mathcal{K}_s$ and $x^m_{\bar{k}nt}, k\in\mathcal{K}_m$ to indicate channel-UE-slot assignments, and binary z_{nt} to represent if a RU is used.

p^s_{knt} SBS's power for UE $k\in\mathcal{K}_s$ on n and t

$p^m_{\bar{k}nt}$ MBS's power for UE $\bar{k}\in\mathcal{K}_m$ on n and t

$$x^s_{knt} = \begin{cases} 1 & \text{if UE } k\in\mathcal{K}_s \text{ is allocated on } n \text{ and } t,\\ 0 & \text{otherwise.}\end{cases}$$

$$x^m_{\bar{k}nt} = \begin{cases} 1 & \text{if } \bar{k}\in\mathcal{K}_m \text{ is allocated on } n \text{ and } t,\\ 0 & \text{otherwise.}\end{cases}$$

$$z_{nt} = \begin{cases} 1 & \text{if channel } n \text{ and slot } t \text{ (one RU) is used,}\\ 0 & \text{otherwise.}\end{cases}$$

In P0, the objective (3a) is to minimize the total transmit power in a frame. Constraints (3b)-(3c) ensure that all the UEs' file requests and proactive files must be delivered. In constraints (3d)-(3e), the total transmit power at the SBS and the MBS cannot exceed their maximum power limit on each time slot t. Constraints (3f) confine the number of allocated UEs on a RU is no more than L. Constraints (3g) state that any UE from $\mathcal{K}_s$ and from $\mathcal{K}_m$ will not be allocated to the same RU, due to the HD operation at the SBS and the avoidance of strong interference. Constraints (3h) show that if any UE is allocated to channel n and slot t, the corresponding indicator $z_{nt}=1$, otherwise zero, where M can be set as a large value, e.g., the total number of RUs. In (3i) the total number of the used RUs is no more than a certain amount $\bar{M}$. The motivation is to limit the resource consumption for the current data-transmission task and let the system have enough spare resources to deal with the dynamic traffic demand from other terminals.

$$\text{P0:}\min_{\substack{p^s_{knt},p^m_{\bar{k}nt}\\ z_{nt},x^s_{knt},x^m_{\bar{k}nt}}} \sum_{n\in\mathcal{N}}\sum_{t\in\mathcal{T}}\Big(\sum_{k\in\mathcal{K}_s} p^s_{knt} + \sum_{\bar{k}\in\mathcal{K}_m} p^m_{\bar{k}nt}\Big) \tag{3a}$$

$$\text{s.t.}\quad \sum_{n\in\mathcal{N}}\sum_{t\in\mathcal{T}} r^s_{knt} \geq d_k,\ \forall k\in\mathcal{K}_s \tag{3b}$$

$$\sum_{n\in\mathcal{N}}\sum_{t\in\mathcal{T}} r^m_{\bar{k}nt} \geq d_{\bar{k}},\ \forall \bar{k}\in\mathcal{K}_m \tag{3c}$$

$$\sum_{k\in\mathcal{K}_s}\sum_{n\in\mathcal{N}} p^s_{knt} \leq P^s_{max},\ \forall t\in\mathcal{T} \tag{3d}$$

$$\sum_{\bar{k}\in\mathcal{K}_m}\sum_{n\in\mathcal{N}} p^m_{\bar{k}nt} \leq P^m_{max},\ \forall t\in\mathcal{T} \tag{3e}$$

$$\sum_{k\in\mathcal{K}_s} x^s_{knt} + \sum_{\bar{k}\in\mathcal{K}_m} x^m_{\bar{k}nt} \leq L,\ \forall n\in\mathcal{N},\ \forall t\in\mathcal{T} \tag{3f}$$

$$x^s_{knt} + x^m_{\bar{k}nt} \leq 1,\ \forall n\in\mathcal{N},\ \forall t\in\mathcal{T}, \forall k\in\mathcal{K}_s, \forall \bar{k}\in\mathcal{K}_m \tag{3g}$$

$$\sum_{k\in\mathcal{K}_s} x^s_{knt} + \sum_{\bar{k}\in\mathcal{K}_m} x^m_{\bar{k}nt} \leq M z_{nt}, \forall n\in\mathcal{N}, \forall t\in\mathcal{T} \tag{3h}$$

$$\sum_{n\in\mathcal{N}}\sum_{t\in\mathcal{T}} z_{nt} \leq \bar{M} \tag{3i}$$

$$p^s_{knt} \leq P^s_{max} x^s_{knt},\ \forall n\in\mathcal{N},\ \forall t\in\mathcal{T}, \forall k\in\mathcal{K}_s \tag{3j}$$

$$p^m_{\bar{k}nt} \leq P^m_{max} x^m_{\bar{k}nt},\ \forall n\in\mathcal{N},\ \forall t\in\mathcal{T}, \forall \bar{k}\in\mathcal{K}_m \tag{3k}$$

Constraints (3j)-(3k) connect the continuous p-variables and the binary x-variables. In the formulation, each p-variable, e.g., p^s_{knt}, is associated with an x-variable, e.g., x^s_{knt}. If a p-variable is positive, the corresponding x-variable is restricted to one. When a p-variable is zero, the x-variable can be zero or one in the constraints. However, the x-variable will be forced to be zero, instead of one, in the optimization process. This is because setting any $p^s_{knt}=0$ but $x^s_{knt}=1$ means that a RU is used but no contribution for delivering UE k's data ($p^s_{knt}=0$ and $r^s_{knt}=0$), which is clearly not optimal.

P0 presents a mixed integer non-linear programming problem. The non-linearity lies in functions $r^s_{knt} = B\log\left(1+\mathrm{SINR}^s_{knt}\right), k\in\mathcal{K}_s$ and $r^m_{\bar{k}nt} = B\log\left(1+\mathrm{SINR}^m_{\bar{k}nt}\right), \bar{k}\in\mathcal{K}_m$. For solving P0, a nec-

essary step is to identify the convexity/non-convexity of the problem when the integer constraint is relaxed. Although P0 is straightforward to describe the considered optimization problem, the convexity/non-convexity may not be easy to identify from the current form. We then reformulate the problem in P1 by the principle of disciplined convex programming in order to make the convexity/non-convexity easy to observe [11]. By performing successive variables substitution in rate functions r^s_{knt} and $r^m_{\bar{k}nt}$, the power variables, p^s_{knt} and $p^m_{\bar{k}nt}$ in P0, can be equivalently expressed by functions of rates [12]. In P1, we treat r^s_{knt} and $r^m_{\bar{k}nt}$ as the optimization variables. The non-linear objective (4a) and constraints (4b), (4c) are convex functions [11]. In (4d) and (4e), analogous to constraints (3j)-(3k), we introduce two large-value parameters R^s_{max} and R^m_{max} to connect r-variables and x-variables.

$$\text{P1:} \min_{\substack{r^s_{knt}, r^m_{\bar{k}nt} \\ z_{nt}, x^s_{knt}, x^m_{\bar{k}nt}}} \sum_{t\in\mathcal{T}}\sum_{n\in\mathcal{N}}\Big(\sum_{k=1}^{|\mathcal{K}_s|}\Big(\frac{\sigma^2}{|h^s_{kn}|^2}-\frac{\sigma^2}{|h^s_{k-1,n}|^2}\Big)e^{\left(\sum_{i=k}^{|\mathcal{K}_s|} r^s_{int}\right)} \tag{4a}$$

$$+\sum_{\bar{k}=0}^{|\mathcal{K}_m|}\Big(\frac{\sigma^2}{|h^m_{\bar{k}n}|^2}-\frac{\sigma^2}{|h^m_{\bar{k}-1,n}|^2}\Big)e^{\left(\sum_{j=\bar{k}}^{|\mathcal{K}_m|} r^m_{jnt}\right)}\Big)$$

$$\text{s.t. } (3b), (3c)$$

$$\sum_{n\in\mathcal{N}}\sum_{k=1}^{|\mathcal{K}_s|}\Big(\frac{\sigma^2}{|h^s_{kn}|^2}-\frac{\sigma^2}{|h^s_{k-1,n}|^2}\Big)e^{\left(\sum_{i=k}^{|\mathcal{K}_s|} r^s_{int}\right)}\le P^s_{max}, \forall t\in\mathcal{T} \tag{4b}$$

$$\sum_{n\in\mathcal{N}}\sum_{\bar{k}=0}^{|\mathcal{K}_m|}\Big(\frac{\sigma^2}{|h^m_{\bar{k}n}|^2}-\frac{\sigma^2}{|h^m_{\bar{k}-1,n}|^2}\Big)e^{\left(\sum_{j=\bar{k}}^{|\mathcal{K}_m|} r^m_{jnt}\right)}\le P^m_{max}, \forall t\in\mathcal{T} \tag{4c}$$

$$(3f), (3g), (3h), (3i)$$

$$r^s_{knt}\le R^s_{max}x^s_{knt},\ \forall n\in\mathcal{N},\ \forall t\in\mathcal{T}, \forall k\in\mathcal{K}_s \tag{4d}$$

$$r^m_{\bar{k}nt}\le R^m_{max}x^m_{\bar{k}nt},\ \forall n\in\mathcal{N},\ \forall t\in\mathcal{T}, \forall \bar{k}\in\mathcal{K}_m \tag{4e}$$

IV. Proposed Optimal Algorithm

A. Solution Characterizations

P1 is a mixed-integer program with a set of linear constraints and non-linear convex constraints. To obtain the optimal solution, the computational complexity is in general high due to the combinatorial nature of the problem. The problem falls into the domain of mixed-integer convex programming (MICP) [13]. A generic MICP can be stated as,

$$\min_{\boldsymbol{y}} f_0(\boldsymbol{y}) \tag{5a}$$

$$\text{s.t. } \boldsymbol{y}\in\boldsymbol{Y},\ \boldsymbol{Y}=\{\boldsymbol{y}: f_i(\boldsymbol{y})\le 0, i=1,\dots,m\} \tag{5b}$$

$$y_j\in\mathbb{Z},\ j=1,\dots,J \tag{5c}$$

The objective $f_0(\boldsymbol{y})$ is linear in a standard form. The convex region $\boldsymbol{Y}$ is formed by the intersection of convex constraint functions $f_1(\boldsymbol{y})\dots, f_m(\boldsymbol{y})$, and $y_j\in\mathbb{Z},\ j=1,\dots,J$ defines integrality constraints. More specifically, P1 is a mixed-integer exponential conic optimization problem (MIECP) where the non-linear convex constraints are in the format of exponential cones [13]. The MIECP belongs to a subset of MICP, and is challenging to solve in general. Some state-of-the-art solvers, e.g., MOSEK (version 8), can be applied to solve MIECP, but only for the instances of $symmetric$ cones, e.g., quadratic cones [14]. The exponential cone is a $non\text{-}symmetric$ cone [11]. Compared to symmetric cones, the solutions for mixed-integer non-symmetric cone optimization are less studied and less mature.

In P1, one can observe that when the integer variables are relaxed, the relaxation problem becomes a continuous convex optimization problem. The branch-and-bound algorithm is a straightforward solution to enable global optimum, where a convex relaxation problem is solved at each node of the branch-and-bound tree. As investigated in the literature, this method is not competitive in computational efficiency for solving many MICP problems [14]. Driven by the powerful mix-integer linear programming (MILP) solver, it is often efficient if one can avoid solving the non-linear convex relaxation problem but solving a polyhedral relaxation problem by MILP [13]. In this paper, we consider an optimal solution based on polyhedral outer approximation (POA). The convex region $\boldsymbol{Y}$ is encompassed and approximated by a polyhedron $\boldsymbol{Q}$, i.e., $\boldsymbol{Y}\subseteq\boldsymbol{Q}$. Set $\boldsymbol{Q}$ is formed by an intersection of a finite number of closed half-spaces, namely linear inequality constraints.

B. Proposed POA Algorithm for Optimally Solving P1

In POA, a lower bound v_{lb} for (5) is derived by relaxing the convex region $\boldsymbol{Y}$ to $\boldsymbol{Q}$ and solving the relaxation problem (6) by MILP.

$$v_{lb}=\min_{\boldsymbol{y}} f_0(\boldsymbol{y}) \tag{6a}$$

$$\text{s.t. } \boldsymbol{y}\in\boldsymbol{Q} \tag{6b}$$

$$y_j\in\mathbb{Z},\ j=1,\dots,J \tag{6c}$$

The integer assignments $y^*_j\in\mathbb{Z},\ j=1,\dots,J$ are obtained by solving (6). By fixing the integer variables with these integer values in (5), an upper bound v_{ub} is obtained by solving the following continuous convex problem (7) with the original convex region $\boldsymbol{Y}$,

$$v_{ub}=\min_{\boldsymbol{y}} f_0(\boldsymbol{y}) \tag{7a}$$

$$\text{s.t. } \boldsymbol{y}\in\boldsymbol{Y} \tag{7b}$$

$$y_j=y^*_j,\ j=1,\dots,J \tag{7c}$$

The POA algorithm performs in a loop between (6) and (7). The algorithm terminates when $v_{ub}=v_{lb}$. If the gap of v_{ub} and v_{lb} is larger than a tolerance ϵ, the algorithm tightens the polyhedral $\boldsymbol{Q}$ by adding constraints (8) which are generated by the first-order approximation,

$$f_i(\boldsymbol{y}^*)+\triangledown f_i(\boldsymbol{y}^*)^T(\boldsymbol{y}-\boldsymbol{y}^*)\le 0, i=1,\dots,m \tag{8}$$

where $\boldsymbol{y}^*$ is the optimal solution of (7). Then (6) with a new $\boldsymbol{Q}$ is resolved to obtain a non-decreasing sequence of lower bounds [15].

In order to apply POA for solving P1, we introduce an auxiliary variable q to enable a linear objective. P1 is then equivalently reformulated as P1$'$. The constraints only containing continuous variables, i.e., (3b), (3c), (4b), (4c), and (9b), form the convex set $\boldsymbol{Y}$, whereas (3f) - (3i), (4d), (4e) represent discrete constraints.

$$\text{P1}': \min_{\substack{q,r^s_{knt},r^m_{\bar{k}nt}\\ z_{nt},x^s_{knt},x^m_{\bar{k}nt}}} \quad q \tag{9a}$$

$$\text{s.t. } (3b),(3c),(4b),(4c),(3f)-(3i),(4d),(4e)$$

$$\sum_{t\in\mathcal{T}}\sum_{n\in\mathcal{N}}\Big(\sum_{k=1}^{|\mathcal{K}_s|}\Big(\frac{\sigma^2}{|h^s_{kn}|^2}-\frac{\sigma^2}{|h^s_{k-1,n}|^2}\Big)e^{\left(\sum_{i=k}^{|\mathcal{K}_s|} r^s_{int}\right)} + \sum_{\bar{k}=0}^{|\mathcal{K}_m|}\Big(\frac{\sigma^2}{|h^m_{\bar{k}n}|^2}-\frac{\sigma^2}{|h^m_{\bar{k}-1,n}|^2}\Big)e^{\left(\sum_{j=\bar{k}}^{|\mathcal{K}_m|} r^m_{jnt}\right)}\Big) \le q \tag{9b}$$

Algorithm 1 Polyhedral Outer Approximation for Solving P1

1: **Initialize**: tolerance ϵ, v_{lb}, v_{ub}, and polyhedron $\boldsymbol{Q}$
2: **while** $v_{ub}-v_{lb}\ge\epsilon$ **do**
3: Update lower bound v_{lb} by solving the MILP problem (10)
4: **if** (10) is infeasible **then**
5: P1 is infeasible, and the algorithm terminates **end**
6: Obtain optimal integer solutions $z^*_{nt}\in\boldsymbol{z}^*$ and $x^{s^*}_{knt}, x^{m^*}_{\bar{k}nt}\in\boldsymbol{x}^*$ from (10)
7: Fix x-variables by $\boldsymbol{x}^*$ and z-variables by $\boldsymbol{z}^*$ in (11)
8: Update upper bound v_{ub} by solving the continuous convex problem (11)
9: Obtain optimal continuous solutions q^* and $r^{s^*}_{knt}, r^{m^*}_{\bar{k}nt}\in\boldsymbol{r}^*$ from (11)
10: Use q^* and $\boldsymbol{r}^*$ to update polyhedron $\boldsymbol{Q}$ according to (8)
end while

The procedure of POA for solving P1 is summarized in Algorithm 1. In the initialization step, an initial polyhedron $\boldsymbol{Q}$ is firstly constructed to encompass $\boldsymbol{Y}$. The polyhedral relaxation problem for P1$'$ is shown in (10), where $\boldsymbol{r}$ is a collection of all the r-variables. Note that constraints (3b), (3c), (4b), (4c), and (9b) are replaced by set $\boldsymbol{Q}$ in (10).

$$\min_{q,r^s_{knt},r^m_{\bar{k}nt},z_{nt},x^s_{knt},x^m_{\bar{k}nt}} q,\ \text{s.t. } \boldsymbol{r}\in\boldsymbol{Q};\ (3f)-(3i),(4d),(4e) \tag{10}$$

Optimally solving (10) in Line 6 provides integer solutions $\boldsymbol{z}^*$ and $\boldsymbol{x}^*$ for z-variables and x-variables, respectively. Fixing the integer variables by $\boldsymbol{z}^*$ and $\boldsymbol{x}^*$ in P1$'$, the resulting convex problem is shown in (11). Note that with the fixed integer solution, constraints (3f) - (3i) are redundant thus are removed from (11).

$$\min_{q,r^s_{knt},r^m_{\bar{k}nt}} q \tag{11a}$$

$$\text{s.t. } (3b),(3c),(4b),(4c),(9b)$$

$$r^s_{knt}\le R^s_{max}x^{s^*}_{knt},\ \forall n\in\mathcal{N},\ \forall t\in\mathcal{T},\forall k\in\mathcal{K}_s \tag{11b}$$

$$r^m_{\bar{k}nt}\le R^m_{max}x^{m^*}_{\bar{k}nt},\ \forall n\in\mathcal{N},\ \forall t\in\mathcal{T},\forall \bar{k}\in\mathcal{K}_m \tag{11c}$$

Optimally solving (11) outputs continuous solution q^* and $\boldsymbol{r}^*$ in Line 9. Since the strong duality holds at (11) and $q^*,\boldsymbol{r}^*$ are optimal for (11), the first-order approximation approach (8) is sufficient to guarantee a finite convergence of the algorithm [14]. We remark that at the convergence, Algorithm 1 provides global optimal solutions (with tolerance ϵ) for the MICP problem [13], [14].

V. Performance Evaluation

In this section, we use the optimal solutions to illustrate the performance of NOMA-enabled cache replacement and content delivery. As a performance benchmark, the OMA-based scheme is evaluated by setting parameter $L=1$ in constraints (3f). The simulation parameters are summarized in Table I. For content updating in the cache, the LRU policy [3] is adopted to determine which files need to be replaced by the proactive caching files $\mathcal{F}_{ms}$.

Table I
Simulation Parameters

Parameter	**Value**
Channel bandwidth, B	1 MHz
Number of channels, N	5
Number of users, K	20
Number of time slots, T	10
Fading	Quasi-static block fading
Noise power spectral density	-173 dBm/Hz
P^m_{max}	43 dBm
P^s_{max}	30 dBm
Tolerance ϵ in Algorithm 1	10^{-4}
Files' popularity distribution	Zipf, skewness factor 0.8
Number of files, F	200
Cache capacity	$0.3F$
Parameter L	2
Caching replacement algorithm	LRU

Firstly, given UEs' requested files $\mathcal{F}_s,\mathcal{F}_m$ and proactive files $\mathcal{F}_{ms}$, Fig. 2 evaluates the total power consumption per frame. The performance is averaged over 100 snapshots. The results demonstrate that for delivering the same amount of data with $\bar{M}$ available RUs, the NOMA-based solution leads to much lower power consumption than OMA. In Fig. 2, we also observe that the NOMA-enabled data delivery is more favorable to the scenarios with limited RU resources. The NOMA-based transmission scheme is able to consume fewer RUs than the OMA-based scheme. This is evident from that once the number of available RUs ($\bar{M}$) becomes smaller than $K=20$, the

OMA-based scheme is immediately infeasible, whereas the NOMA-based scheme can still complete the data-transmission task but consuming more transmit power.

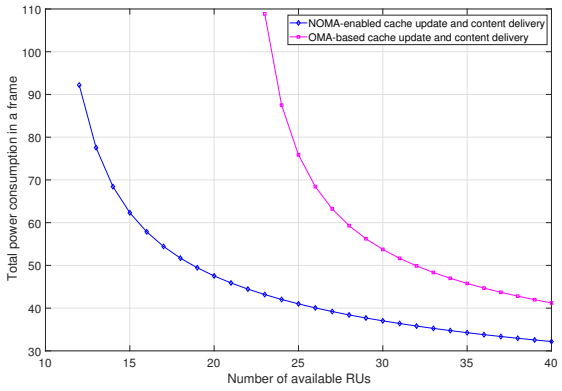

Figure 2. The average total power consumption (W) per frame with respect to the number of available RUs, $\bar{M}$.

Next, we show the long-term performance comparison between frequent and infrequent cache replacement. The simulation is carried out over 10^4 frames. UEs' file requests $\mathcal{F}_u$ are updated frame-by-frame. Algorithm 1 is performed once per frame. The cache replacement is executed every 10 and 100 frames for the frequently-update scheme, and every 5000 frames for the infrequently-update scheme. NOMA is applied to enable efficient cache update along with UEs' data delivery. Compared to the infrequently-update scheme, although more data in $\mathcal{F}_{ms}$ needs to be transmitted in the frequently-update schemes, the total power consumption is on average lower. This is because by timely updating and proactively pushing files to the cache, the cache hit probability increases in subsequent frames. As a result, more UEs will be served by the local cache at the SBS with low power consumption, instead of requesting MBS data from the cell edge. On the other hand, one can also observe that too frequent cache updating, e.g., every 10 frames, may not necessarily bring significant performance gains due to excessive power consumed for frequently transmitting the proactive caching files. Therefore a trade-off between the cache-update frequency and the power consumption exists an needs to be optimized.

Table II
AVERAGE TOTAL POWER PER FRAME WITH RESPECT TO VARIOUS CACHE-UPDATE PERIOD

Cache replacement period	Avg. total power per frame (W)
Every 10 frames	115.42
Every 100 frames	71.35
Every 5000 frames	128.58

VI. Conclusions

We considered applying NOMA-based data transmission to enable efficient cache replacement and content delivery. We formulated a min-power resource allocation problem for updating cache's content and satisfying users' data demand. The considered problem is identified as a mixed-integer exponential conic optimization problem. We designed a polyhedral outer approximation algorithm to obtain the globe optimal solution. Numerical results demonstrate the promising performance of the considered NOMA-enabled data transmission, in terms of power and the capability of completing data delivery in resource-limited scenarios.

VII. Acknowledgments

The work has been supported by the European Research Council (ERC) project AGNOSTIC, and by the FNR CORE projects ROSETTA (11632107) and ProCAST.

References

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

[2] L. Lei, L. You, G. Dai, T. X. Vu, D. Yuan, and S. Chatzinotas, "A deep learning approach for optimizing content delivering in cache-enabled HetNet," in Proc. *IEEE ISWCS*, Aug. 2017.

[3] A. Gharaibeh, I. Hababeh and M. Alshawaqfeh, "An Efficient online cache replacement algorithm for 5G networks," in *IEEE Access*, vol. 6, pp. 41179–41187, 2018.

[4] T. X. Vu, L. Lei, S. Vuppala, A. Kalantari, S. Chatzinotas, and B. Ottersten, "Latency minimization for content delivery networks with wireless edge caching", in Proc. *IEEE ICC*, May 2018.

[5] L. Lei, D. Yuan, C. K. Ho, and S. Sun, "Optimal cell clustering and activation for energy saving in load-coupled wireless networks," in *IEEE Transactions on Wireless Communications*, vol. 14, no. 11, pp. 6150–6163, Nov. 2015.

[6] Z. Ding, P. Fan, G. K. Karagiannidis, R. Schober and H. V. Poor, "NOMA assisted wireless caching: strategies and performance analysis," in *IEEE Transactions on Communications*, vol. 66, no. 10, pp. 4854-4876, Oct. 2018.

[7] Z. Zhao, M. Xu, Y. Li and M. Peng, "A non-orthogonal multiple access-based multicast scheme in wireless content caching networks," in *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2723-2735, Dec. 2017.

[8] L. Lei, L. You, Q. He, T. X. Vu, S. Chatzinotas, D. Yuan, and B. Ottersten, "Learning-assisted optimization for energy-efficient scheduling in deadline-aware NOMA systems," in *IEEE Transactions on Green Communications and Networking*, pre-print, 2019.

[9] L. Xiang, D. W. K. Ng, X. Ge, Z. Ding, V. W. S. Wong and R. Schober, "Cache-aided non-orthogonal multiple access," in Proc. *IEEE ICC*, May 2018.

[10] S. Mehrizi, A. Tsakmalis, S. Chatzinotas, B. Ottersten, "Bayesian learning for content popularity estimation in edge-Caching networks", in Proc. *IEEE CCNC*, Jan. 2019.

[11] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, UK: Cambridge University Press, 2004.

[12] L. Lei, D. Yuan and P. Värbrand, "On power minimization for non-orthogonal multiple access (NOMA)," in *IEEE Communications Letters*, vol. 20, no. 12, pp. 2458-2461, Dec. 2016.

[13] A. Vinel, P. Krokhmal, "Mixed integer programming with a class of nonlinear convex constraints," *Discrete Optimization*, vol. 4, pp. 66–86, 2017.

[14] M. Lubin, E. Yamangil, R. Bent, et al. "Polyhedral approximation in mixed-integer convex optimization," *Mathematical Programming*, vol. 172, pp. 139–168, 2018.

[15] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientific, 1999.