

Distributed Cooperative Caching for VoD with Geographic Constraints

Konstantin Avrachenkov
INRIA
Sophia Antipolis, France
konstantin.avrachenkov@inria.fr

Jasper Goseling
Mathematics of Operations Research
University of Twente, The Netherlands
j.goseling@utwente.nl

Berksan Serbetci*
EURECOM
Sophia Antipolis, France
serbetci@eurecom.fr
*Corresponding author.

Abstract— We consider caching of video streams in a cellular network in which each base station is equipped with a cache. Video streams are partitioned into multiple substreams and the goal is to place substreams in caches such that the residual backhaul load is minimized. We consider two coding mechanisms for the substreams: Layered coding (LC) mechanism and multiple description coding (MDC). We develop a distributed asynchronous algorithm for deciding which files to store in which cache to minimize the residual bandwidth, *i.e.*, the cost for downloading the missing substreams of the user's requested video with a certain video quality from the gateway (*i.e.*, the main server). We show that our algorithm converges rapidly. Finally, we show that MDC partitioning is better than the LC mechanism when the most popular content is stored in caches; however, our algorithm enables to use the LC mechanism as well without any performance loss.

I. INTRODUCTION

Data traffic in cellular networks is rapidly expanding and one of the main reasons is the explosive growth in the demand for Video-on-Demand (VoD) content. Recent studies show that existing network infrastructures will not be able to support the demand [1]. One of the bottlenecks will be formed by the backhaul links that connect base stations to the core network and we need to utilize these links as efficiently as possible. A promising means to efficiently handle this issue is to proactively cache data in the base stations. The idea is to store part of the popular data at the wireless edge and use the backhaul to refresh the stored data at off-peak hours since users' demand distribution over time is varying slowly. In order to embody this, content and network service providers are deploying Video Hub Offices (VHOs) serving as caches in metropolitan areas to serve VoD subscribers in that area. In this way, caches containing popular content serve as helpers to the overall system and decrease the backhaul load.

Our goal in this paper is on developing low-complexity distributed and asynchronous content placement algorithms for VoD. As content and network service providers would like to optimize the cached content in caches (*i.e.*, base stations, or more specifically VHOs) while keeping the communication in the network to a minimum, the problem is of practical relevance in cellular networks. Our low-complexity distributed and

asynchronous algorithm is designed to optimize the content and keeping the communication between caches to a minimum as the caches exchange information only locally.

Let us first provide a brief discussion of the related works. Caching has been receiving a lot of attention in the literature lately. Here we provide a brief overview of the work on caching that is most closely related to the current paper. A general outline of a distributed caching architecture for wireless networks considering a network model in which a bipartite graph indicates how users are connected to base stations has been presented in [14]. Works in [2], [4], [6], [13] consider placement of base stations in the plane according to a stochastic geometry where the main aim is to find the optimal placement strategy for the network for various performance measures. In [5] we have developed a distributed asynchronous algorithm that deals with the miss probability minimization by casting the problem into the framework of potential games for such networks where caches are located at arbitrary locations, and showed that our algorithm converges to a Nash equilibrium, and in fact to the best Nash equilibrium in most practical scenarios. Works in [11], [12] also consider the problem of caching from a game-theoretic perspective for different cost functions. Finally in [3], authors present an approach for optimal content placement for a large-scale VoD by formulating the problem as a mixed integer problem. The main difference between [5], [3] and the present work is the considered performance measures, the concept of video partitioning and the heterogeneity of the video popularities over the network. For a more thorough review of the existing caching work, the reader is strongly encouraged to refer to papers [5], [10].

In the remainder of the introduction we will give an overview of the model and contributions.

We consider a network with caches located at arbitrary locations in the plane. Caches know their own coverage area as well as the coverage areas of other caches that overlap with this region. There is a content catalog from which users request videos with a certain video quality according to a known probability distribution. Videos are partitioned into chunks (*i.e.*, substreams) by using two coding mechanisms: Layered coding (LC) mechanism [17] and Multiple Description Coding (MDC) [7]. Each cache can store a limited number of chunks and the goal is to minimize the residual bandwidth, *i.e.*, the

This work was supported by the European Research Council under the EU Horizon 2020 research and innovation program / ERC grant agreement no. 725929.

cost for downloading the missing video chunks of the users' requested videos with certain video qualities from the main server. We develop low-complexity asynchronous distributed cooperative content placement caching algorithms that require communication only between caches with overlapping coverage areas. In the basic algorithm, at each iteration a cache will selfishly update its cache content by minimizing the local residual bandwidth and by considering the content stored by neighbouring caches. We show that our basic algorithm can be formulated as providing the best response dynamics in a potential game. We show that our algorithm is guaranteed to converge to a Nash equilibrium in finite time. We analyze the structure of the best response dynamics and provide the closed-form optimal solutions for the content placement by solving the linear optimization problem. Finally, we illustrate our results by a number of numerical results with various video quality request distributions for real world network models.

To specify, our contributions are as follows:

- We provide the partitioning models for the two aforementioned coding mechanisms and formulate problems in order to minimize the residual bandwidth for both mechanisms;
- We provide a distributed asynchronous algorithm for optimizing the content placement which can be interpreted as giving the best response dynamics in a potential game for both partitioning mechanisms;
- We prove that the best response dynamics can be obtained as a solution of a linear optimization problem;
- We prove that our algorithm converges in finite time, and in fact rapidly;
- We evaluate our algorithm through numerical examples using the base station locations from a real wireless network for the cellular network topology. We study the residual bandwidth evolution on a real network numerically for both coding mechanisms and show that our distributed caching algorithm strongly suggests using the LC mechanism for video partitioning in VoD systems;

Let us outline the organization of the paper. In Section II we give the formal model and problem definitions. In Section III we provide the game formulation of the problem, analyze the structures of the best response dynamics and Nash equilibria. In Section IV, we present practical implementations of our low-complexity algorithms for various video quality distributions and show the resulting performance for a real network. In Section V we conclude the paper with some discussions and provide an outlook on future research.

II. MODEL AND PROBLEM DEFINITION

We consider a network of N base stations that are located in the plane \mathbb{R}^2 . We will use the notation $[1 : N] = \{1, \dots, N\}$ and $\Theta = \mathbb{P}([1 : N]) \setminus \emptyset$, where $\mathbb{P}([1 : N])$ is the power set of $[1 : N]$. We specify the geometric configuration of the network through A_s , $s \in \Theta$, which denotes the area of the plane that is covered only by the caches in subset s , namely $A_s = (\cap_{\ell \in s} \bar{A}_\ell) \cap (\cap_{\ell \notin s} \bar{A}_\ell^c)$, where \bar{A}_ℓ is the complete coverage region of cache ℓ .

As a special case we will consider the case that all base stations have same circular coverage region with radius r . In this case we specify the location of each base station, with x_m for the location of base station $m \in [1 : N]$. We then obtain \bar{A}_m as the disc of radius r around x_m .

The content library consists of videos from the set $\mathcal{C} = \{c_1, \dots, c_J\}$, where $J < \infty$. Each element c_j represents a video, where j is the file index. The size of video c_j is denoted by w_j . Next, videos are partitioned into chunks. We assume that video c_j is partitioned into Q chunks and consequently, the chunk library consists of video chunks $\mathcal{C}_c = \{c_{1,1}, \dots, c_{1,Q}, \dots, c_{J,1}, \dots, c_{J,Q}\}$, where $J, Q < \infty$. Each element $c_{j,q}$ represents a video chunk, where j is the file index and q is the chunk index. The chunk size for the element $c_{j,q}$ is denoted by $w_{j,q}$. Each base station is equipped with a cache that can be used to store video chunks from \mathcal{C}_c .

The motivation behind partitioning videos into chunks is as follows. In this work we will consider various coding mechanisms. If the video is partitioned by using layered coding (LC) mechanisms (e.g., MPEG-2, MPEG-4, etc.), there is a base layer and enhancement layers. The base layer is necessary for the media stream to be decoded. Accordingly, the further enhancement layers are applied to improve the video quality. Hence, for the layered coding mechanism, the element $c_{j,1}$ represents the base layer for video j , and $c_{j,2}$ is the first enhancement layer required to obtain a better video quality, and so forth. For the LC mechanism, $Q = Q_{LC}$ and the chunk size $w_{j,q}$ will then depend on the average base size of the different video qualities, and the total size of the video j with the highest video quality is equal to

$$w_j = \sum_{q=1}^{Q_{LC}} w_{j,q}. \quad (1)$$

Another coding mechanism that we will consider is multiple description coding (MDC). For MDC, the video stream is partitioned into multiple substreams, referred to as descriptions. Any description can be used to decode the media stream. As the number of received substreams increases, the received video quality increases. Since there is no hierarchical layered coding system, each received unique description is equally beneficial to increase the video quality. For the MDC mechanism, $Q = Q_{MDC}$ and the chunk size $w_{j,q}$ is constant for any chunk q for fixed video j , and the total size of the video j with the highest video quality is equal to

$$w_j = Q_{MDC} w_{j,q}. \quad (2)$$

Our interest is in users located in the plane over the area that is covered by the base stations, *i.e.*, uniformly distributed in $A_{cov} = \cup_{s \in \Theta} A_s$. The probability of a user in the plane being covered by caches $s \in \Theta$ (and is not covered by additional caches) is denoted by $p_s = |A_s|/|A_{cov}|$. A user located in A_s , $s \in \Theta$ can connect to all caches in subset s , and has access to all video chunks stored in these caches.

We assume that there are R video quality levels. The probability that video j is requested with video quality ρ by

a user at $x_u \in s$ is denoted by $a_{j,\rho,s}$. Here $\rho = 1$ refers to the video with the lowest quality and $\rho = R$ refers to the case where the video has the highest quality. There is an obvious relation between the required chunks and the desired video quality. For the LC mechanism, files can be divided into $Q_{LC} = R$ layers with the proper size selections ($w_{j,q}$) and receiving all layers gives the video with the highest quality. On the other hand, for the MDC mechanism, in practice Q_{MDC} is not equal to R since all chunks have equal size. We define a new parameter D_ρ that represents the required number of chunks to get the video quality ρ .

We assume that the distribution for the requested video's quality depends on the user's location. To validate this assumption, one might think that users travelling in the bus will more likely request watching videos in a lower resolution compared to those located in residential areas. Similarly, users located in a residential area may request watching videos in different qualities as some users may watch videos via their ultra high-definition televisions, whereas others watch videos via their mobile phones with a lower resolution. We assume that the requested video quality for a user located in $s \in \Theta$ follows a known and fixed probability mass function (pmf) $f^{(s)}(\rho)$.

For the file popularities, we assume that $a_1 \geq a_2 \geq \dots \geq a_J$ for any video quality ρ . We assume that the video popularities do not vary as ρ changes. Even though any popularity distribution can be used, most of our numerical results will be based on the Zipf distribution for the video popularities for any video quality ρ . For any video quality indicator ρ , the probability that a user will ask for content j is equal to

$$a_j = \frac{j^{-\gamma}}{\sum_{j=1}^J j^{-\gamma}}, \quad (3)$$

where $\gamma > 0$ is the Zipf parameter.

Since a_j and $f^{(s)}(\rho)$ represent statistically independent random variables, we conclude that the probability that video j is requested by a user located at $x_u \in s$ with video quality ρ is equal to

$$a_{j,\rho,s} = \frac{j^{-\gamma}}{\sum_{j=1}^J j^{-\gamma}} f^{(s)}(\rho). \quad (4)$$

Content is placed in caches using knowledge of the request statistics $a_{j,\rho,s}$, but without knowing the actual request made by the user. We denote the placement policy for cache m as

$$b_{j,q}^{(m)} := \begin{cases} 1, & \text{if } c_{j,q} \text{ is stored in cache } m, \\ 0, & \text{if } c_{j,q} \text{ is not stored in cache } m, \end{cases} \quad (5)$$

and the overall placement strategy for cache m as

$$\mathbf{B}^{(m)} = \begin{bmatrix} b_{1,1}^{(m)} & \dots & b_{J,1}^{(m)} \\ \vdots & \ddots & \vdots \\ b_{1,Q}^{(m)} & \dots & b_{J,Q}^{(m)} \end{bmatrix}$$

as a $Q \times J$ matrix. The overall placement strategy for the network is denoted by $\mathbf{B} = [\mathbf{B}^{(1)}; \dots; \mathbf{B}^{(N)}]$ as an $Q \times J \times N$ three-dimensional matrix.

Caches have capacity K , i.e.,

$$\sum_{j=1}^J \sum_{q=1}^Q w_{j,q} b_{j,q}^{(m)} \leq K, \forall m.$$

For clarity of presentation, we assume homogeneous capacity for the caches. However, our work can immediately be extended to the network topologies where caches have different capacities.

The user requests to watch one of the videos from the content library with certain quality. The aim is to place content in the caches ahead of time in order to minimize the residual bandwidth, i.e., the cost for downloading the missing video chunks of the user's requested video with a certain video quality from the gateway in order to serve the users. Next, we will design cache placement strategies for residual bandwidth minimization for the LC and MDC mechanisms.

A. Layered coding mechanism

For the LC mechanism, the residual bandwidth $f_{LC}(\mathbf{B})$ is equal to

$$f_{LC}(\mathbf{B}) = \sum_{s \in \Theta} \sum_{j=1}^J \sum_{\rho=1}^R p_s a_{j,\rho,s} \left[\sum_{q=1}^{\rho} w_{j,q} \prod_{\ell \in s} (1 - b_{j,q}^{(\ell)}) \right]. \quad (6)$$

Our goal is to find the optimal placement strategy minimizing the residual bandwidth as follows:

Problem 1.

$$\begin{aligned} \min \quad & f_{LC}(\mathbf{B}) \\ \text{s.t.} \quad & \sum_{j=1}^J \sum_{q=1}^Q w_{j,q} b_{j,q}^{(m)} \leq K, \forall m, \\ & b_{j,q}^{(m)} \in \{0, 1\}, \forall j, q, m. \end{aligned} \quad (7)$$

It is easy to verify that Problem 1 is not convex. We will provide a distributed asynchronous algorithm to address Problem 1 in which we iteratively update the placement policy at each cache. In [5] we have showed for a different performance measure that we can define an algorithm that can be viewed as the best response dynamics in a potential game. We will present a similar algorithm here. We make use of the following notation. Denote by $\mathbf{B}^{(-m)}$ the placement policies of all caches except cache m . We will write $f_{LC}(\mathbf{B}^{(m)}, \mathbf{B}^{(-m)})$ to denote $f_{LC}(\mathbf{B})$. Also, for the sake of simplicity in the following section, let $f_{LC}^{(m)}$ denote the residual bandwidth of cache m , i.e.,

$$\begin{aligned} f_{LC}^{(m)}(\mathbf{B}) &= \sum_{\substack{s \in \Theta \\ m \in s}} \sum_{j=1}^J \sum_{\rho=1}^R \sum_{q=1}^{\rho} p_s a_{j,\rho,s} w_{j,q} (1 - b_{j,q}^{(m)}) \\ &\quad \prod_{\ell \in s \setminus \{m\}} (1 - b_{j,q}^{(\ell)}) \\ &= \sum_{\substack{s \in \Theta \\ m \in s}} \sum_{j=1}^J \sum_{\rho=1}^R \sum_{q=1}^{\rho} p_s a_{j,\rho,s} w_{j,q} (1 - b_{j,q}^{(m)}) \zeta^{(m)}(j, q), \end{aligned} \quad (9)$$

where

$$\zeta^{(m)}(j, q) = \prod_{\ell \in s \setminus \{m\}} (1 - b_{j,q}^{(\ell)}). \quad (10)$$

B. Multiple description coding

In MDC, receiving all Q_{MDC} descriptions will provide the highest expected video quality experience for the users. For the sake of simplicity, we define a new parameter $y_{j,\rho,s}$ denoting the total number of descriptions that needs to be fetched from the backhaul in order to satisfy the expected video quality ρ for video j in $s \in \Theta$ as

$$y_{j,\rho,s} = \max \left\{ 0, D_\rho - Q + \sum_{q=1}^Q \prod_{\ell \in s} (1 - b_{j,q}^{(\ell)}) \right\}. \quad (11)$$

For the MDC mechanism, the residual bandwidth $f_{MDC}(\mathbf{B})$, i.e., the cost for downloading the missing chunks of the user's requested video from the gateway is equal to

$$f_{MDC}(\mathbf{B}) = \frac{1}{Q} \sum_{s \in \Theta} \sum_{j=1}^J \sum_{\rho=1}^R p_s a_{j,\rho,s} w_j y_{j,\rho,s}. \quad (12)$$

Our goal is to find the optimal placement strategy minimizing the total number of descriptions that needs to be fetched from the backhaul, equivalently minimizing the residual bandwidth as follows:

Problem 2.

$$\begin{aligned} \min & f_{MDC}(\mathbf{B}) \\ \text{s.t.} & \sum_{j=1}^J \sum_{q=1}^Q w_j b_{j,q}^{(m)} \leq KQ, \quad \forall m, \end{aligned} \quad (13)$$

$$b_{j,q}^{(m)} \in \{0, 1\}, \quad \forall j, q, m, . \quad (14)$$

As in the LC mechanism, Problem 2 is not convex either. Therefore, we will provide a distributed asynchronous algorithm to address Problem 2 (similar to the procedure we followed for the layered mechanism) in which we iteratively update the placement policy at each cache. We will make use of the following notation. Denote by $\mathbf{B}^{(-m)}$ the placement policies of all caches except cache m . We will write $f_{MDC}(\mathbf{B}^{(m)}, \mathbf{B}^{(-m)})$ to denote $f_{MDC}(\mathbf{B})$. Also, for the sake of simplicity for the potential game formulation that will be presented in the following section, let $f_{MDC}^{(m)}$ denote the residual bandwidth of cache m , i.e.,

$$\begin{aligned} f_{MDC}^{(m)}(\mathbf{B}) &= \frac{1}{Q} \sum_{\substack{s \in \Theta \\ m \in s}} \sum_{j=1}^J \sum_{\rho=1}^R p_s a_{j,\rho,s} w_j \\ &\max \left\{ 0, D_\rho - Q + \sum_{q=1}^Q (1 - b_{j,q}^{(m)}) \zeta^{(m)}(j, q) \right\}, \end{aligned} \quad (15)$$

where $\zeta^{(m)}(j, q)$ is given in (10).

III. POTENTIAL GAME FORMULATION

In this section we provide a distributed asynchronous algorithm to address Problems 1 and 2 in which we iteratively update the placement policy at each cache. We will show that this algorithm can be formulated as providing the best response dynamics in a potential game.

In our algorithm, each cache tries selfishly to minimize its residual bandwidth $f_{LC}^{(m)}$ and $f_{MDC}^{(m)}$ defined in (9) and (15) for the LC and MDC mechanisms respectively. Given a placement $\mathbf{B}^{(-m)}$ by the other caches, cache m solves for $\mathbf{B}^{(m)}$ in

Problem 3.

$$\begin{aligned} \min & f_{LC}^{(m)}(\mathbf{B}^{(m)}, \mathbf{B}^{(-m)}) \\ \text{s.t.} & \sum_{j=1}^J \sum_{q=1}^Q w_{j,q} b_{j,q}^{(m)} \leq K, \\ & b_{j,q}^{(m)} \in [0, 1], \quad \forall j, q, \end{aligned} \quad (16)$$

for the LC mechanism and,

Problem 4.

$$\begin{aligned} \min & f_{MDC}^{(m)}(\mathbf{B}^{(m)}, \mathbf{B}^{(-m)}) \\ \text{s.t.} & \sum_{j=1}^J \sum_{q=1}^Q w_j b_{j,q}^{(m)} \leq KQ, \quad \forall m, \\ & b_{j,q}^{(m)} \in [0, 1], \quad \forall j, q, \end{aligned} \quad (18)$$

$$b_{j,q}^{(m)} \in [0, 1], \quad \forall j, q, \quad (19)$$

for the MDC mechanism.

Each cache continues to optimize its placement strategy until no further improvements can be made. At this point \mathbf{B} is a *Nash equilibrium strategy* that minimizes the overall residual bandwidth satisfying

$$f_{LC}^{(m)}(\mathbf{B}^{(m)}, \mathbf{B}^{(-m)}) \leq f_{LC}^{(m)}(\bar{\mathbf{B}}^{(m)}, \mathbf{B}^{(-m)}), \quad \forall m, \bar{\mathbf{B}}^{(m)}, \quad (20)$$

for the LC mechanism and,

$$f_{MDC}^{(m)}(\mathbf{B}^{(m)}, \mathbf{B}^{(-m)}) \leq f_{MDC}^{(m)}(\bar{\mathbf{B}}^{(m)}, \mathbf{B}^{(-m)}), \quad \forall m, \bar{\mathbf{B}}^{(m)}, \quad (21)$$

for the MDC mechanism.

For both coding mechanisms, we will refer to these games as the *content placement games* and demonstrate in the next subsections that both games are potential games [9] with many nice properties.

A. Convergence analysis

In this section, we will prove for both coding mechanisms that if the caches repeatedly update their placement strategies it is guaranteed to converge to a Nash equilibrium in finite time. Note that the order of the updates is not important as long as all caches are scheduled infinitely often.

Theorem 1. *The content placement games defined by payoff functions (9) and (15) are potential games with the potential function given in (6) and (12), respectively. For both games,*

if we schedule each cache infinitely often, the best response dynamics converges to a Nash equilibrium in finite time.

Proof. In order to show that both games are potential with the potential functions $f_{LC}(\mathbf{B})$ and $f_{MDC}(\mathbf{B})$, it is easy to verify that

$$\begin{aligned} & f_{LC}^{(m)}(\bar{\mathbf{B}}^{(m)}, \mathbf{B}^{(-m)}) - f_{LC}^{(m)}(\mathbf{B}^{(m)}, \mathbf{B}^{(-m)}) \\ &= f_{LC}(\bar{\mathbf{B}}^{(m)}, \mathbf{B}^{(-m)}) - f_{LC}(\mathbf{B}^{(m)}, \mathbf{B}^{(-m)}), \end{aligned}$$

and

$$\begin{aligned} & f_{MDC}^{(m)}(\bar{\mathbf{B}}^{(m)}, \mathbf{B}^{(-m)}) - f_{MDC}^{(m)}(\mathbf{B}^{(m)}, \mathbf{B}^{(-m)}) \\ &= f_{MDC}(\bar{\mathbf{B}}^{(m)}, \mathbf{B}^{(-m)}) - f_{MDC}(\mathbf{B}^{(m)}, \mathbf{B}^{(-m)}), \end{aligned}$$

which completes the proof of the first statement. The proof states that the improvement in the residual bandwidth by the best response after each update is equal to the improvement in the residual bandwidth in the overall network. The detailed analysis is trivial and skipped due to space constraints.

Now, since there exists only a finite number of placement strategies, none of the caches will be missed in the long-run. Moreover, each non-trivial best response provides a positive improvement in the potential function in a potential game. Hence, we are guaranteed to converge to a Nash equilibrium in finite time for both games. \square

B. Structure of the best response dynamics

In this subsection we will analyze the structure of the best response dynamics. We will show that solutions to Problem 3 and to Problem 4 can be obtained by solving a linear optimization problem and we provide the solutions in closed form.

For notational purposes, let us first define the set $\Omega = \{(j, q), (j \in [1, J]), (q \in [1 : Q])\}$ that consists of all video index-video quality tuples.

Theorem 2. *The optimal solution to Problem 3 is given by*

$$\bar{b}_{j,q}^{(m)} = \begin{cases} 1, & \text{if } \pi_m^{-1}(j, q) \leq K/w_{j,q}, \\ 0, & \text{if } \pi_m^{-1}(j, q) > K/w_{j,q}, \end{cases} \quad (22)$$

where $\pi_m : \Omega \rightarrow [1 : J \times Q]$ satisfies

$$\begin{aligned} & \sum_{\substack{s \in \Theta \\ m \in s}} \sum_{\rho = \pi_m(1,2)}^R p_s a_{\pi_m(1,1), \rho, s} w_{\pi_m(1,1), \pi_m(1,2)} \\ & \quad \zeta^{(m)}(\pi_m(1,1), \pi_m(1,2)) \\ & \geq \sum_{\substack{s \in \Theta \\ m \in s}} \sum_{\rho = \pi_m(2,2)}^R p_s a_{\pi_m(2,1), \rho, s} w_{\pi_m(2,1), \pi_m(2,2)} \\ & \quad \zeta^{(m)}(\pi_m(2,1), \pi_m(2,2)) \geq \dots \\ & \geq \sum_{\substack{s \in \Theta \\ m \in s}} \sum_{\rho = \pi_m(J \times Q, 2)}^R p_s a_{\pi_m(J \times Q, 1), \rho, s} w_{\pi_m(J \times Q, 1), \pi_m(J \times Q, 2)} \\ & \quad \zeta^{(m)}(\pi_m(J \times Q, 1), \pi_m(J \times Q, 2)), \end{aligned}$$

where $\pi_m(i, 1)$ is the file index j of the i^{th} sorted tuple, and $\pi_m(i, 2)$ is the video quality index q of the i^{th} sorted tuple.

Proof. The objective function $f_{LC}^{(m)}(\mathbf{B})$ is an affine function and we have several inequality constraints. It is easy to verify that we also have affine functions in the constraints. Therefore, we can conclude that Problem 3 is a linear problem. It can then be easily solved using the dual problem. We have Lagrange dual function

$$\begin{aligned} & \inf_{\mathbf{B}} \left(\sum_{\substack{s \in \Theta \\ m \in s}} \sum_{j=1}^J \sum_{\rho=1}^R \sum_{q=1}^Q p_s a_{j, \rho, s} w_{j, q} (1 - b_{j, q}^{(m)}) \zeta^{(m)}(j, q) + \right. \\ & \quad \lambda \left(\sum_{j=1}^J \sum_{q=1}^Q w_{j, q} b_{j, q}^{(m)} - K \right) - \sum_{j=1}^J \sum_{q=1}^Q \mu_{j, q} b_{j, q}^{(m)} + \\ & \quad \left. \sum_{j=1}^J \sum_{q=1}^Q \nu_{j, q} (b_{j, q}^{(m)} - 1) \right). \end{aligned}$$

For notational convenience we introduce the functions $\Psi_{j,q} : \mathbb{R} \rightarrow [0, 1]$, $j = 1, \dots, J$, and $q = 1, \dots, Q$ as follows

$$\Psi_{j,q}(\lambda) = \begin{cases} 1, & \text{if } \lambda < \sum_{\substack{s \in \Theta \\ m \in s}} \sum_{\rho=q}^R p_s a_{j, \rho, s} w_{j, q} \zeta^{(m)}(j, q), \\ 0, & \text{if } \lambda \geq \sum_{\substack{s \in \Theta \\ m \in s}} \sum_{\rho=q}^R p_s a_{j, \rho, s} w_{j, q} \zeta^{(m)}(j, q). \end{cases} \quad (23)$$

We also define $\Psi : \mathbb{R} \rightarrow [0, K]$, where $\Psi(\lambda) = \sum_{j=1}^J \sum_{q=1}^Q w_{j, q} \Psi_{j, q}(\lambda) = K$ for

$$\lambda \in \left(-\infty, \sum_{\substack{s \in \Theta \\ m \in s}} \sum_{\rho=q}^R p_s a_{j, \rho, s} w_{j, q} \zeta^{(m)}(j, q) \right),$$

and $\Psi(\lambda) = \sum_{j=1}^J \sum_{q=1}^Q w_{j, q} \Psi_{j, q}(\lambda) = 0$ for

$$\lambda \in \left[\sum_{\substack{s \in \Theta \\ m \in s}} \sum_{\rho=q}^R p_s a_{j, \rho, s} w_{j, q} \zeta^{(m)}(j, q), \infty \right).$$

It is then possible to check all possible combinations from the video content-video quality tuples $\Omega \rightarrow [1 : J \times Q]$ to confirm if the condition given in (23) is satisfied. In order to satisfy the capacity constraint (16) the above solution is guaranteed to exist. The proof is completed by validating that with the strategy above, $\bar{\lambda}$ solving the Lagrange dual is satisfying $\Psi(\bar{\lambda}) = K$. \square

The intuition behind Theorem 2 is as follows. The ordering is simply made based on the weights of the tuples consisting of the probability of being located in a specific region, the summation coming from the effect of the difference between requests with different video qualities (and the summation index difference due to the effect of the layered coding mechanism), and the effect of the possibility of receiving the chunks from the neighbours.

Theorem 3. *The optimal solution to Problem 4 is given by*

$$\bar{b}_{j,q}^{(m)} = \begin{cases} 1, & \text{if } \pi_m^{-1}(j, q) \leq KQ/w_j, \\ 0, & \text{if } \pi_m^{-1}(j, q) > KQ/w_j, \end{cases} \quad (24)$$

where $\pi_m : \Omega \rightarrow [1 : J \times Q]$ satisfies

$$\begin{aligned}
& \sum_{\substack{s \in \Theta \\ m \in s}} \sum_{\rho=1}^R p_s a_{\pi_m(1,1),\rho,s} w_{\pi_m(1,1)} \zeta^{(m)}(\pi_m(1,1), \pi_m(1,2)) \\
& \quad \mathbb{1}(D_\rho - Q + \sum_{i=1}^Q \zeta^{(m)}(\pi_m(1,1), i) > 0) \\
& \geq \sum_{\substack{s \in \Theta \\ m \in s}} \sum_{\rho=1}^R p_s a_{\pi_m(2,1),\rho,s} w_{\pi_m(2,1)} \zeta^{(m)}(\pi_m(2,1), \pi_m(2,2)) \\
& \quad \mathbb{1}(D_\rho - Q + \sum_{i=1}^Q \zeta^{(m)}(\pi_m(2,1), i) > 0) \geq \dots \\
& \geq \sum_{\substack{s \in \Theta \\ m \in s}} \sum_{\rho=1}^R p_s a_{\pi_m(J \times Q, 1), \rho, s} w_{\pi_m(J \times Q, 1)} \zeta^{(m)}(\pi_m(J \times Q, 1), \\
& \quad \pi_m(J \times Q, 2)) \mathbb{1}(D_\rho - Q + \sum_{i=1}^Q \zeta^{(m)}(\pi_m(J \times Q, 1), i) > 0),
\end{aligned}$$

where $\pi_m(i, 1)$ is the file index j of the i^{th} sorted tuple, and $\pi_m(i, 2)$ is the video quality index q of the i^{th} sorted tuple., and $\mathbb{1}(\cdot)$ is the indicator function, i.e., it is equal to 1 if the condition is satisfied and 0 otherwise.

Proof. The objective function $f_{MDC}^{(m)}(\mathbf{B})$ is a linear function and we have several inequality constraints. It is easy to verify that we also have linear functions in the constraints. Therefore, we can conclude that Problem 4 is a linear problem. It can then be easily solved using the dual problem. We have Lagrange dual function

$$\begin{aligned}
& \inf_{\mathbf{B}} \left(\sum_{\substack{s \in \Theta \\ m \in s}} \sum_{j=1}^J \sum_{\rho=1}^R p_s a_{j,\rho,s} w_j \right. \\
& \quad \left. \max \left\{ 0, D_\rho - Q + \sum_{q=1}^Q \left(1 - b_{j,q}^{(m)} \right) \zeta^{(m)}(j, q) \right\} + \right. \\
& \quad \left. \lambda \left(\sum_{j=1}^J \sum_{q=1}^Q w_j b_{j,q}^{(m)} - KQ \right) - \sum_{j=1}^J \sum_{q=1}^Q \mu_{j,q} b_{j,q}^{(m)} + \right. \\
& \quad \left. \sum_{j=1}^J \sum_{q=1}^Q \nu_{j,q} \left(b_{j,q}^{(m)} - 1 \right) \right).
\end{aligned}$$

For notational convenience we introduce the functions $\Phi_{j,q} : \mathbb{R} \rightarrow [0, 1]$, $j = 1, \dots, J$, and $q = 1, \dots, Q$ as follows

$$\Phi_{j,q}(\lambda) = \begin{cases} 1, & \text{if } \lambda < \sum_{\substack{s \in \Theta \\ m \in s}} \sum_{\rho=q}^R p_s a_{j,\rho,s} \zeta^{(m)}(j, q), \\ 0, & \text{if } \lambda \geq \sum_{\substack{s \in \Theta \\ m \in s}} \sum_{\rho=q}^R p_s a_{j,\rho,s} \zeta^{(m)}(j, q), \end{cases} \quad (25)$$

when $D_\rho - Q + \sum_{i=1}^Q \zeta^{(m)}(j, i) > 0$. We also define $\Phi : \mathbb{R} \rightarrow [0, K]$, where $\Phi(\lambda) = \sum_{j=1}^J \sum_{q=1}^Q w_j \Phi_{j,q}(\lambda) / Q = K$ for

$$\lambda \in \left(-\infty, \sum_{\substack{s \in \Theta \\ m \in s}} \sum_{\rho=q}^R p_s a_{j,\rho,s} w_{j,q} \zeta^{(m)}(j, q) \right),$$

and $\Phi(\lambda) = \sum_{j=1}^J \sum_{q=1}^Q w_j \Phi_{j,q}(\lambda) = 0$ for

$$\lambda \in \left[\sum_{\substack{s \in \Theta \\ m \in s}} \sum_{\rho=q}^R p_s a_{j,\rho,s} w_{j,q} \zeta^{(m)}(j, q), \infty \right).$$

It is then possible to check all possible combinations from the video content-video quality tuples $\Omega \rightarrow [1 : J \times Q]$ to confirm if the condition given in (25) is satisfied. In order to satisfy the capacity constraint (18) the above solution is guaranteed to exist. The proof is completed by validating that with the strategy above, $\bar{\lambda}$ solving the Lagrange dual is satisfying $\Phi(\bar{\lambda}) = K$. \square

C. Structure of Nash equilibria

In this subsection we provide insight into the structure of the Nash equilibria of the content placement games. We know from the previous subsection that both games are potential games. Hence, the Nash equilibria for the games correspond to the optimal placement strategies satisfy the solutions of the dual problems of Problem 1 and Problem 2, respectively.

Corollary 1. *Let $\bar{\mathbf{B}}$ denote a placement strategy at a Nash equilibrium of the content placement game for the LC mechanism. Then, $\bar{\mathbf{B}}$ satisfies (22) $\forall m = 1, \dots, N$.*

Proof. The proof is similar to the proof of Theorem 2, where in this case the optimal placement strategies must hold for all $1 \leq m \leq N$ simultaneously. \square

Corollary 2. *Let $\bar{\mathbf{B}}$ denote a placement strategy at a Nash equilibrium of the content placement game for the MDC mechanism. Then, $\bar{\mathbf{B}}$ satisfies (24) $\forall m = 1, \dots, N$.*

Proof. The proof is similar to the proof of Theorem 3, where in this case the optimal placement strategies must hold for all $1 \leq m \leq N$ simultaneously. \square

IV. NUMERICAL EVALUATION

In this section we will present practical implementations of our algorithm for the content placement games for the LC and MDC mechanisms and evaluate our theoretical results according to a network of caches with their geographical locations following a real wireless network.

A. The ROBR algorithm

We will use the Random Order Best Response (ROBR) algorithm [5] for the content placement games for both coding mechanisms to minimize the residual bandwidth. The basic idea of our algorithm is to repeatedly perform best response dynamics presented in Section III-B. For ROBR algorithm, at each iteration step, a random cache is chosen

uniformly from the set $[1 : N]$ and updated by applying best response dynamics. We assume that all caches are initially empty (i.e., $\mathbf{B}^{(m)} = 0_{Q,J}$, where $0_{Q,J}$ represents a $Q \times J$ zero matrix, $\forall m \in [1 : N]$). The algorithm stops when $f_{LC}^{(m)}(\mathbf{B}^{(m)}, \mathbf{B}^{(-m)})$ converges for the LC mechanism, and $f_{MDC}^{(m)}(\mathbf{B}^{(m)}, \mathbf{B}^{(-m)})$ converges for the MDC mechanism. ROBR algorithm is shown in Algorithm 1.

Algorithm 1: Random Order Best Response (ROBR)

```

initialize  $\mathbf{B}^{(m)} = 0_{Q,J}, \forall m \in [1 : N]$ ;
set  $imp(m) = 1, \forall m \in [1 : N]$ ;
set  $\mathbf{imp} = [imp(1), \dots, imp(N)]$ ;
while  $\mathbf{imp} \neq \mathbf{0}$  do
     $m = \text{Uniform}(N)$ ;
    Set  $imp(m) = 0$ ;
    Solve Problem 3* for cache  $m$  and find  $\bar{\mathbf{B}}^{(m)}$  using
        the information coming from neighbours;
    Compute  $f_{LC}^{(m)}(\bar{\mathbf{B}}^{(m)}, \mathbf{B}^{(-m)})$ *;
    if  $f_{LC}^{(m)}(\bar{\mathbf{B}}^{(m)}, \mathbf{B}^{(-m)}) - f_{LC}^{(m)}(\mathbf{B}^{(m)}, \mathbf{B}^{(-m)}) \neq 0$ *
        then
             $imp(m) = 1$ 
        end
    end
end

```

*For the MDC mechanism, the algorithm solves Problem 4 and uses the performance measure $f_{MDC}^{(m)}(\bar{\mathbf{B}}^{(m)}, \mathbf{B}^{(-m)})$.

B. A real wireless network: Berlin network

In this section we will evaluate our theoretical results for the topology of a real wireless network. We have taken the positions of the base stations provided by the OpenMobileNetwork project [18]. The base stations are located in the area 1.95×1.74 kms around the TU-Berlin campus. We will consider these base stations as our caches with certain cache capacities. The coverage radius of the base stations is equal to $r = 700$ m. The locations of the base stations is shown in Figure 1.

We consider the content library of size $J = 200$. We assume a Zipf distribution for the video popularities, setting $\gamma = 1$ and taking a_j according to (3). We consider the case where videos

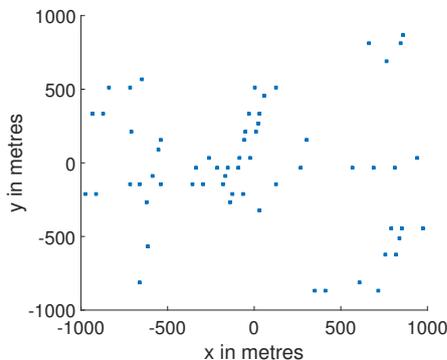


Fig. 1. Locations of Base Stations from the OpenMobileNetwork dataset.

TABLE I
VIDEO QUALITY SPECIFICATIONS

Video quality (ρ)	Base Size (MB/min)	Video Size (GB)
1 (240p)	2.56	0.1024
2 (360p)	4.30	0.1720
3 (480p)	6.79	0.2716
4 (720p)	15.49	0.6196
5 (1080p)	32.70	1.3080

have $R = 5$ different video qualities. We use the video quality bandwidth requirements and the base size data given in [8] for the chunk sizes, where the base size is the average of the sizes of a large set of tracked videos in certain video qualities given in megabytes per minute (MB/min). We assume that videos in the content library are all 40 minutes long. The corresponding video quality specifications are shown in Table I. The first column shows the video quality index ρ , the second column shows the base sizes of the videos and the third column shows the 40 minutes long video sizes with different qualities.

The videos are partitioned into $Q_{LC} = 5$ chunks by the LC mechanism. From the third column of Table I, it immediately follows that $w_{j,1} = 102.4$ MB, $w_{j,2} = 69.6$ MB, $w_{j,3} = 99.6$ MB, $w_{j,4} = 348.0$ MB, and $w_{j,5} = 688.4$ MB, $\forall j$. By the MDC mechanism, the videos are partitioned into $Q_{MDC} = 38$ chunks and $w_{j,q} = 34.42$ MB $\forall j, q$. Consequently, $D_1 = 3$, $D_2 = 5$, $D_3 = 8$, $D_4 = 18$, and $D_5 = 38$. Note that the selection of Q_{MDC} is based on the least common multiple of the third column of Table I. Finally, we set $K = 6.54$ GB, i.e., each cache can store five 1080p movies.

For the distribution of the requested video qualities, we have

$$f^{(s)}(\rho) = \begin{cases} \frac{1+t_\rho\theta_s}{5+10\theta_s}, & \text{if } \rho \in 1, 2, 3, 4, 5, \\ 0, & \text{otherwise,} \end{cases} \quad (26)$$

where t_ρ is the parameter representing the direction of the increment of the requested video qualities and θ_s is the slope of the linear increment.

We will now evaluate our theoretical results for three different video quality trends.

1) *Uniform*: For this trend, we set $\theta_s = 0, \forall s \in \Theta$. This trend reflects the case where users request videos in all qualities equally likely throughout the whole network.

2) *Linear correlation*: For this trend, we set $t_\rho = 5 - \rho$ for $|s| < \mathbb{E}[|s|]$, and $t_\rho = \rho - 1$ for $|s| > \mathbb{E}[|s|]$, and $\theta_s = 0.1 ||s| - \mathbb{E}[|s|]|$, where $|s|$ is the cardinality of the subset $s \in \Theta$, $\mathbb{E}[|s|]$ is the mean of the cardinalities of the all subsets, and $||s| - \mathbb{E}[|s|]|$ is the absolute value of the difference of them. This trend reflects the case where users covered by many caches request videos in higher quality.

3) *Inverse correlation*: We use the same parameters as in linear correlation case with the reversed t_ρ 's, i.e., $t_\rho = \rho - 1$ for $|s| < \mathbb{E}[|s|]$, and $t_\rho = 5 - \rho$ for $|s| > \mathbb{E}[|s|]$. This trend reflects the case where users covered by many caches request videos in lower quality.

In Figure 2 the residual bandwidth evolution for three different video quality trends for the proposed algorithm for both coding mechanisms is shown. In all video quality trends,

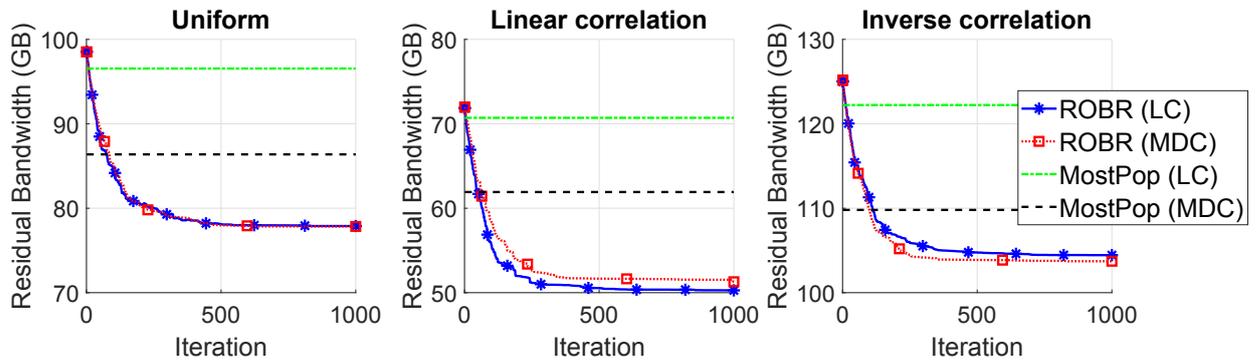


Fig. 2. The residual bandwidth evolution.

storing the most popular content for the MDC mechanism gives a lower residual bandwidth compared to the LC mechanism since chunks have identical size and any received chunk increases the video quality for MDC. However, after running ROBR, the advantage of using MDC vanishes as both coding mechanisms give the same residual bandwidth performance. One might easily think that MDC would perform better than LC for a VoD system since any received chunk would increase the video quality. However, we see that LC can perform as good as MDC by applying ROBR and placing the chunks optimally. This is a crucial observation due to following reason. From a technical point of view, at a given bitrate budget, there is a quality loss in MDC with respect to LC and chunks are more difficult to encode [15]. In fact [16] shows that some serious portion of the bitrate budget is reserved for forward error correction in MDC and the actual residual bandwidth will be higher for MDC to obtain the exact same quality. Finally, the residual bandwidth performance is the best for the linear correlation case as the requested video quality increases as the number of caches that users can connect to increases.

V. DISCUSSION AND CONCLUSION

In this paper we have provided a low-complexity asynchronously distributed cooperative caching algorithm for VoD when there is communication only between caches with overlapping coverage areas. We have related our algorithm to a best response dynamics in a game and have shown with practical examples that it converges rapidly. We have demonstrated the residual bandwidth evolution on a real network and have shown that our algorithm diminishes the advantage of using MDC over the LC mechanism. Due to technical details, MDC suffers from quality loss since the chunks are more difficult to encode. Hence, we conclude that using LC for video partitioning is better for a VoD system as long as ROBR is applied for the optimal placement of the video chunks.

In future work we will look into this problem from the users' perspective and focus on maximizing the user utilities.

REFERENCES

- [1] "Cisco VNI global mobile data traffic forecast, 2017-2022," San Jose, CA, USA, February 2019.
- [2] E. Altman, K. Avrachenkov, and J. Goseling, "Distributed storage in the plane," *Networking Conference, IFIP 2014*, pp. 1–9, Trondheim, Norway, June 2014.
- [3] D. Applegate, A. Archer, V. Gopalakrishnan, S. Lee, and K. K. Ramakrishnan, "Optimal content placement for a large-scale VoD system," *IEEE/ACM Transactions on Networking*, vol. 24, no. 4, pp. 2114–2127, August 2016.
- [4] K. Avrachenkov, X. Bai, and J. Goseling, "Optimization of caching devices with geometric constraints," in press for *International Journal of Performance Evaluation*, also *ArXiv: 1602.03635*, 2016.
- [5] K. Avrachenkov, J. Goseling, and B. Serbetci, "A low-complexity approach to distributed cooperative caching with geographic constraints," *Proceedings of the ACM on Measurement and Analysis of Computing Systems (POMACS)*, vol. 1, no. 1, pp. 1–25, June 2017.
- [6] B. Blaszczyszyn, and A. Giovanidis, "Optimal geographic caching in cellular networks," *IEEE International Conference on Communications (ICC) 2015*, pp. 3358–3363, London, UK, June 2015.
- [7] V. K. Goyal, "Multiple description coding: compression meets the network," *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 74–93, September 2001.
- [8] D. K. Krishnappa, D. Bhat, and M. Zink, "DASHing YouTube: An analysis of using DASH in YouTube video service," *38th Annual IEEE Conference on Local Computer Networks*, pp. 407–415, Sydney, Australia, October 2013.
- [9] D. Monderer, and L. Shapley, "Potential games," *Games and Economic Behavior*, vol. 14, pp. 124–143, 1996.
- [10] G. S. Paschos, G. Iosifidis, M. Tao, D. Towsley, and G. Caire, "The Role of caching in future communication systems and networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 6, pp. 1111–1125, June 2018.
- [11] F. De Pellegrini, A. Massaro, L. Goratti, and R. El-Azouzi, "A pricing scheme for content caching in 5G mobile edge clouds," *IEEE WINCOM 2016*, Fez, Morocco, 2016.
- [12] F. De Pellegrini, A. Massaro, L. Goratti, and R. El-Azouzi, "Competitive caching of contents in 5G edge cloud networks," *IEEE WiOpt CCDWN 2017*, Paris, France, 2017.
- [13] B. Serbetci, and J. Goseling, "Optimal geographical caching in heterogeneous cellular networks," *arXiv: 1710.09626*, 2017.
- [14] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: wireless content delivery through distributed caching helpers," *IEEE Transactions on Information Theory*, vol. 59, no.12, pp. 8402–8413, December 2013.
- [15] A. Vitali, "Multiple description coding - a new technology for video streaming over the Internet," *EBU Technical Review*, October 2007.
- [16] Y. Wang, A. R. Reibman, and S. Lin, "Multiple description coding for video delivery," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 57–70, January 2005.
- [17] T. Wiegand, G.J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, July 2003.
- [18] OpenMobileNetwork, <http://map.openmobilenetwork.org/>.