Transmission Delay Minimization via Joint Power Control and Caching in Wireless HetNets

Derya Malak[*], Faruk V. Mutlu[†], Jinkun Zhang[†], and Edmund M. Yeh[†]
[*]Communication Systems Dept., EURECOM, 06904 Biot Sophia Antipolis cedex, FRANCE, derya.malak@eurecom.fr
[†]Northeastern University, Boston, MA, USA, {mutlu.f, zhang.jinku}@northeastern.edu, eyeh@ece.neu.edu

***Abstract*—A fundamental challenge in wireless heterogeneous networks (HetNets) is to effectively use the limited transmission and storage resources in the presence of increasing deployment density and backhaul capacity constraints. To alleviate bottlenecks and reduce resource consumption, we design optimal caching and power control algorithms for multi-hop wireless HetNets. We devise a joint optimization framework to minimize the average transmission delay as a function of the caching variables and the signal-to-interference-plus-noise ratios (SINR) as determined by the transmission powers, while explicitly accounting for backhaul connection costs and the power constraints.**

Using convex relaxation and rounding, we obtain a reduced-complexity formulation (RCF) of the joint optimization problem, which can provide a constant factor approximation to the globally optimal solution. We characterize the necessary (KKT) conditions for an optimal solution to RCF, and use strict quasi-convexity to show that the KKT points are Pareto optimal for RCF. We then devise a subgradient projection algorithm to jointly update the caching and power variables, and show that under appropriate conditions, the algorithm converges at a linear rate to the local minima of RCF, under general SINR. We support our analytical findings with results from numerical experiments.

***Index Terms*—Joint power-caching optimization, biconvexity, quasi-convexity, Pareto optimality, subgradient algorithm.**

I. Introduction

The energy and cost efficiencies of wireless heterogeneous networks (HetNets) incorporating macro cells (MCs) and small cells (SCs) are critical for meeting the performance requirements of 5G wireless networks [1]. Design of these HetNets entails the fundamental challenge of optimally utilizing both the bandwidth and storage resources of the network to reduce the download or transmission delay and the energy costs. With the increasing deployment density in wireless networks, the backhaul capacity becomes the bottleneck. It is well known that caching can alleviate this bottleneck by replacing the backhaul capacity with storage capacity at SCs [2], i.e., moving content closer to the wireless edge. Caching reduces transmission delay by bringing the popular data items in SCs that are faster or computationally cheaper to access than MCs. To optimize resource usage in wireless HetNets, designing caching and power control policies and the interplay between caching and transmission decisions remains an open challenge. Enabling this will help control the interference and minimize the transmission delay costs in wireless HetNet topologies.

A. Current State of the Art and Motivation

Research to date on cost optimization of caching has focused on different perspectives. There have been attempts to devise replacement algorithms that aim to optimize the caching gain, which is the reduction in the expected total download delay achieved by caching at intermediate nodes. Simple, elegant, adaptive, and distributed approaches determining how to populate caches in a variety of networking applications abound. These include Che's analytical approximation to compute the probability of an item being in a Least Recently Used (LRU) cache [3], and extension of Che's decoupling approach for a unified analysis for different replacement policies in [4].

Recently, information centric networking (ICN) architectures have put emphasis on the traffic engineering and caching problems [5] to effectively use both bandwidth and storage for efficient content distribution and optimize the network performance [6]. Alternatively, there have been works focusing on jointly optimizing the caching gain and resource usage, e.g., local caching and broadcasting as characterized in the landmark paper [7], and a decentralized SC caching optimization, i.e., femtocaching, to minimize the download delay [2], distributed optimization of caching gain given routing [5], jointly optimizing caching and routing to provide latency guarantees by taking into account congestion [8], and elastic and inelastic traffic [9]. Existing strategies have also focused on separately optimizing the caching gain or the throughput [10], and optimizing spatial throughput via scheduling [11]. From a resource management perspective, it is not sufficient to exclusively optimize caching or throughput, or delay.

There exist several pertinent power control algorithms to optimize the resource usage in wireless networks [12], or maximize throughput under latency considerations [13]. However, delay optimization in wireless links is challenging because of interference and congestion. There exist power-aware routing algorithms for packet forwarding to balance the traffic between high-quality links and less reliable links, such as [14], joint optimization of power control, routing, and congestion [15], and resource optimization under latency and power constraints [16], as well as delay-optimal computation task scheduling at the mobile edge [17], and the minimum delay routing algorithm [18]. In addition, fog optimization-based effective resource allocation schemes for wireless networks have been devised in [19] to achieve high power efficiency and a high Quality of Experience under latency constraints, and in [20] to maximize the sum rate of cellular networks. However, none of these approaches or research on ICN architectures has jointly designed traffic engineering and cache placement strategies to optimize network performance in view of traffic demands.

Despite the advent of different caching solutions, to the best of our knowledge, none of them focuses on the joint

ISBN 978-3-903176-37-9 © 2021 IFIP

optimization of caching and power allocation or provides algorithmic performance guarantees in terms of the achievable costs via caching. Although intermediate caching alleviates the average download delay, it is hard to quantify how this delay is affected by the resource allocation strategy in a HetNet setting. In this paper, we focus on jointly optimizing the network level performance in terms of transmission delay and caching, which can be increasingly skewed away from a strategy that places the items without accounting for the transmission delay.

B. Methodology and Contributions

We study jointly optimal caching and power control for arbitrary multi-hop wireless HetNet topologies with nodes that have caching capabilities. As the networks are becoming increasingly heterogeneous, MCs and SCs can co-exist in 5G and beyond [1]. Dense SC deployment is the key for 5G networks to enhance the capacity, rendering a cost-efficient backhaul solution a key challenge.

For a given caching HetNet topology with multi-hop transmissions, we devise an algorithm for jointly optimal caching and power control to minimize the average transmission delay cost, i.e., the average download delay, per request. While end-to-end delay in systems is due to several key sources, including transmission, propagation, processing and queuing, we are primarily interested in a lightly loaded regime for which congestion-dependent latency costs can be neglected, and each node can sustain a high service rate relative to the average rate at which items are arriving to be serviced. To accurately determine the transmission delay, we explicitly account for the transmission power, backhaul costs, and wireless interference.

Finding the optimum placement of files is proven to be NP-complete [2]. Hence, jointly optimal power control and caching to minimize the transmission delay is also NP-complete. Our joint optimization framework is significantly different from the traditional approach which maximizes the caching gain only. This approach has been widely studied in the literature, such as in [2], [5], [21] and their follow-up works, where the link costs are fixed. This assumption is only true when the links are granted orthogonal frequencies and do not interfere, and the transmission powers are fixed, which is not the case in HetNets. Furthermore, when link costs are deterministic, caching gain always improves with increasing link costs. This requires high transmission powers and violates the purpose of cost minimization. In other words, savings via intermediate caching do not inform us about the actual achievable delay-cost via caching. This justifies our proposed framework in Sect. III, where we consider the minimum achievable cost via caching by taking into account the joint behavior of link costs under resource constraints.

Our main technical contributions include the following:

- **A reduced-complexity formulation (RCF) to the joint optimization problem.** We provide a constant factor approximation to the minimum average transmission delay-cost $D^o(X,S)$ of serving a request via jointly optimizing binary caching variables X and real valued transmission powers S. Using convex relaxation techniques, we obtain an RCF of the joint optimization problem, with cost function $D(Y,S)$ which is not jointly convex, where Y denote the relaxed caching variables. We then round Y to obtain an integral solution within a constant factor from the optimal solution to $D^o(X,S)$.
- **Sufficient conditions for biconvexity of $D(Y,S)$.** We provide a sufficient condition for the convexity of RCF in the logarithm of powers which yields a biconvex RCF objective. This condition pertains to the high SINR regime and does not hold for general SINR values.
- **Joint optimization framework.** We jointly optimize RCF under the general setting which is not jointly convex. We obtain the following results: **a)** strict quasi-convexity of $D(Y,S)$, **b)** generalized necessary conditions for optimality of $D(Y,S)$ assuming strict convexity of $\mathcal{D}_S$, and **c)** Pareto optimality of the solution to $D(Y,S)$.
- **Subgradient projection algorithm.** We provide a subgradient projection algorithm which is guaranteed to converge to a local minimum of the RCF. Due to the non-differentiability and non-convexity of the relaxed problem, we propose a subgradient projection algorithm with a modified Polyak's step size. We also give a simple method to calculate the projection and show that the algorithm converges at a linear rate.

II. Wireless Caching Model

We consider a multi-hop wireless HetNet topology consisting of different types of nodes, e.g., small cells (SCs), macro cells (MCs), and users. The network serves content requests routed over different paths. We assume that radio range of each node is smaller than network coverage area, and nodes have multi-hop capability. We represent the network as a directed graph $\mathcal{G}(V,E)$ where V is the set of nodes such that a node $v\in V$ is either an MC, an SC or a user. All nodes V transmit on the same frequency[1], i.e., all transmissions interfere with each other. In $\mathcal{G}$, E is the set of edges, where given $v,\ u\in V$, the edge $(v,u)\in E$ denotes the transmission link from v to u. In Fig. 1, we illustrate the network and possible multi-hop paths where the users request different items.

The caching model is as follows. The entire set of content items, i.e., the catalog, is denoted by $\mathcal{C}$. Each item in $\mathcal{C}$ is of equal size. Each node is associated with a cache that can store a finite number of content items. The cache capacity at node $v\in V$ is c_v. The variables $x_{vi}\in\{0,1\}$ indicate whether $v\in V$ stores item $i\in\mathcal{C}$. Due to this finite capacity constraint, $\sum_{i\in\mathcal{C}}x_{vi}\leq c_v$, $\forall v\in V$. Each item $i\in\mathcal{C}$ is associated with a fixed set of designated sources $\mathcal{S}_i\subseteq V$, i.e., nodes that always store i: $x_{vi}=1$, $\forall v\in\mathcal{S}_i$. The designated sources could be users, SCs or MCs.

Users issue requests for content items. The set of all requests is denoted by $\mathcal{R}$. A request $r\in\mathcal{R}$ is a pair (i,p) jointly determined by the item $i\in\mathcal{C}$ being requested, and the fixed path p traversed (request is forwarded from the user toward

[1] If subsets of nodes are allocated different frequencies, as in OFDMA-based networks, then we can determine the resulting subset of interfering nodes [22]. This also reduces the interference and improves the coverage performance.

ISBN 978-3-903176-37-9 © 2021 IFIP

a designated source over a fixed path) to serve this request. The routing strategy of a user with respect to $(i, p) \in \mathcal{R}$ is predetermined, e.g., the shortest path in terms of the number of hops to the nearest designated source. We assume that (i) the collection of requests for the same content item i, i.e., $\{p : (i, p) \in \mathcal{R}\}$, are served separately instead of being aggregated, (ii) the response of request (i, p) travels the same path p, in the reverse direction, (iii) different frequency bands are used for the uplink and downlink, (iv) transmission delays are solely due to response messages carrying desired items assuming that request forwarding and cache downloads are instantaneous.

Request rates are known a priori, where choices of requested items are independent. The arrivals of requests are Poisson where the arrival rate of $r = (i, p)$ is $\lambda_{(i,p)}$. A path p on $\mathcal{G}$ of length $|p| = K$ is a sequence $\{p_1, p_2, \ldots, p_K\}$ of nodes $p_k \in V$ such that edge $(p_k, p_{k+1}) \in E$, for $k \in \{1, \ldots, |p| - 1\}$. Let $k_p(v) = \{k \in \{1, \ldots, |p|\} : p_k = v\}$ denote the position of v in p. For each request (i, p), p_1 is the requesting user and $p_{|p|}$ is the designated source of item i, and we assume that p is a simple path, i.e., p contains no loops.

End-to-end delay includes several key components, such as transmission delay, propagation delay, processing delay, and queueing delay. In this paper, we primarily focus on lightly loaded systems, where transmission delay is the dominant one and the other delay components are negligible. We assume there is one queue for each link $(v, u) \in E$ that serves in a first-in-first-out (FIFO) manner all requests traversing (u, v).

To determine the transmission delay of link $(v, u) \in E$ corresponding to (i, p), we first derive the signal-to-interference-plus-noise ratio (SINR) on link (v, u), which we denote by $\mathrm{SINR}_{vu}(S)$, where $S = [s_{vu}] \in \mathbb{R}^{|E|}$ represents the set of transmission powers at all links $(v, u) \in E$. To decode the requests (i, p) traversing link (u, v), we calculate the SINR on link (v, u), where we treat all other transmissions from nodes $j \in V \backslash v$, as well as the transmissions from v to $w \neq u$ as noise. Therefore, the SINR on link (v, u) is given as

$$\mathrm{SINR}_{vu}(S) = \frac{G_{vu} s_{vu}}{N_u + \sum\limits_{j \in V \backslash v} G_{ju} \sum\limits_{w} s_{jw} + G_{vu} \sum\limits_{w \neq u} s_{vw}}, \quad (1)$$

where N_u is the receiver noise power at node u, and s_{vu} is the transmit power from $v \in V$ to u. The total transmit power of node v is $\sum_{u:(v,u)\in E} s_{vu}$. The parameter G_{vu} is the channel power gain that includes only path loss, where we use the standard power loss propagation model, i.e., $G_{vu} = r_{vu}^{-n}$ given distance r_{vu} between v and u, and the path loss exponent $n > 2$. In our model, the transmission delays are *coupled*, in contrast to [2], [21], because the decoding model captures the interference due to simultaneous wireless transmissions. Because the SINR analysis in (1) is for a single frequency band, the set of nodes with nonzero transmission powers causes interference to the unintended receiver node. Employing OFDMA-based schemes allows frequency multiplexing by moving the interfering nodes to orthogonal resources and eliminates the out-of-band interference, and improves the SINR quality. However, we leave this extension to future work.

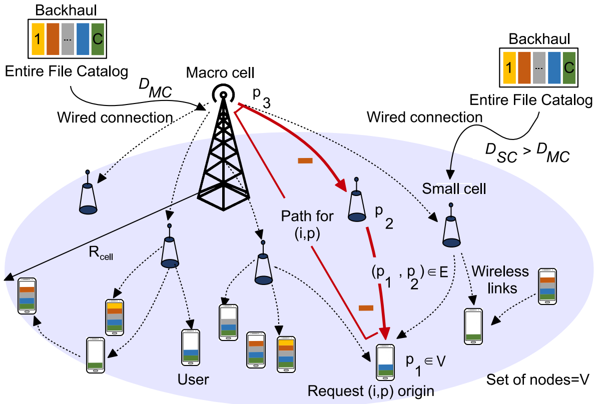

Fig. 1: A caching network scenario with possible connections between the users, SCs or MCs, and to the backhaul. A path $p = \{p_1, p_2, p_3\}$ for request (i, p) is indicated where p_1 is a user, p_2 is a SC, and p_3 is the MC.

To model the wireless transmission delay on link $(v, u) \in E$, we use the following composite relation:

$$f(\mathrm{SINR}_{vu}(S)) = \frac{1}{\log_2(1 + \mathrm{SINR}_{vu}(S))}, \quad (2)$$

which is the delay in number of channel uses per bit corresponding to the data rate of link (v, u). This model captures interference, and thus provides a more sophisticated way of modeling delay in a lightly loaded network than simple hop count. When the SINR is high, (2) yields a low delay and vice versa. From (1)-(2), it is clear that $f(\mathrm{SINR}_{vu}(S))$ is convex and decreasing in $\mathrm{SINR}_{vu}(S)$ but non-convex in S.

Our goal is to jointly optimize the transmission power allocations along with the caching decisions to minimize the average transmission delay of requested items over the multi-hop network. We next formulate this problem.

III. Joint Power Control and Caching Optimization for Transmission Delay Minimization

We formulate the delay minimization problem that jointly considers power control and caching allocations. Due to its NP-hard nature, in Sect. III-A we first develop a RCF based on convex relaxation and its optimal solution, which yields an integral solution (via rounding) whose cost is within a constant factor from that of the optimal solution to the original problem. Next in Sect. III-B we provide a sufficient condition for the convexity of RCF in the logarithm of powers which yields a biconvex objective. This sufficient condition corresponds to the high SINR regime. In Sect. III-C under the general setting which is not jointly convex, we provide various results on the RCF objective. We demonstrate **a)** strict quasi-convexity of $D(Y, S)$, **b)** generalized necessary conditions for optimality of $D(Y, S)$ under strict convexity of $\mathcal{D}_S$, and **c)** Pareto optimality of the solution to $D(Y, S)$. Later in Sect. III-D we provide a subgradient projection algorithm that attains the necessary conditions, along with a linear convergence rate guarantee.

A. Caching Optimization for RCF

A goal in caching systems is to minimize the expected total file downloading delay, i.e., the expected delivery time of content items averaged over the demands and the placement. Since

ISBN 978-3-903176-37-9 © 2021 IFIP

end-to-end delay in our setup is mainly due to the transmission delay, by letting matrix $X = [x_{vi}] \in \{0,1\}^{|V|\times|C|}$ denote the global caching strategy, we can express the cost function for serving a request (i,p) in terms of the transmission delay as

$$D^o_{(i,p)}(X,S) = \sum_{k=1}^{|p|-1} f(\mathrm{SINR}_{p_{k+1}p_k}(S)) \prod_{l=1}^{k}(1 - x_{p_l i}) \ , \quad (3)$$

where $D^o_{(i,p)}(X,S)$ includes the transmission delay of an edge (p_{k+1},p_k) in the path $p = \{p_1,\dots p_k\}$ if none of the nodes $p_1,\dots p_k$ caches i. The last node of p is the designated source, hence a request is always served. Let D^o be the aggregate expected cost in terms of the average number of channel uses per bit, which equals

$$D^o(X,S) = \sum_{(i,p)\in\mathcal{R}} \lambda_{(i,p)} D^o_{(i,p)}(X,S) \ . \quad (4)$$

The gain of intermediate caching is equivalent to the achievable reduction in the overall transmission delay. An upper bound on the expected cost is obtained when all requests are served by the designated sources at the end of each path, i.e.,

$$D^{\mathrm{ub}}(S) = \sum_{(i,p)\in\mathcal{R}} \lambda_{(i,p)} \sum_{k=1}^{|p|-1} f(\mathrm{SINR}_{p_{k+1}p_k}(S)) \ . \quad (5)$$

Our primary objective is to solve the problem

$$\min\{D^o(X,S) : X \in \mathcal{D}_X,\, S \in \mathcal{D}_S\} \ , \quad (6)$$

where $\mathcal{D}_X$ is the feasible set of X satisfying the capacity, integrality, and source constraints:

$$\mathcal{D}_X = \Big\{ \sum\nolimits_{i\in C} x_{vi} \le c_v,\ \forall v \in V,\ x_{vi} \in \{0,1\},\ v \in V,\ i \in \mathcal{C}; \\ x_{vi} = 1,\ \forall i \in \mathcal{C},\ v \in \mathcal{S}_i \Big\}. \quad (7)$$

Letting $O_v = \{u \in V : (v,u) \in E\}$, the feasible set of S is specified by the individual power budget for each node:

$$\mathcal{D}_S = \Big\{ \sum\nolimits_{u\in O_v} s_{vu} \le \hat{s}_v,\ s_{vu} \ge 0,\ \forall v \in V \Big\}, \quad (8)$$

Minimization of $D^o(X,S)$ for a given S subject to $X \in \mathcal{D}_X$ is a reduction from the 2-disjoint set cover problem [2], which is NP-hard. Thus, we devise a centralized algorithm to produce an allocation within a constant approximation of the optimal, without prior knowledge of the topology, edge weights, or the demand distribution. We next formulate a convex relaxation.

a) Convex Relaxation: To approximate the non-convex function $D^o(X,S)$, we construct a convex relaxation, following the approach of [5], [2]. Suppose that x_{vi}, $v \in V$, $i \in \mathcal{C}$, are independent Bernoulli random variables. Let ν be the corresponding joint probability distribution defined over matrices in $\{0,1\}^{|V|\times|C|}$, and denote by $\mathbb{P}_\nu[\cdot]$ and $\mathbb{E}_\nu[\cdot]$ the probability and expectation with respect to ν, respectively.

Relaxing the integrality constraints of X, let marginal probabilities $y_{vi} = \mathbb{P}[x_{vi} = 1] = \mathbb{E}_\nu[x_{vi}] \in [0,1], \forall v \in V, i \in \mathcal{C}$. Denote the feasible set of $Y = [y_{vi}] \in \mathbb{R}^{|V|\times|\mathcal{C}|}$ by $\mathcal{D}_Y = \{\sum_{i\in\mathcal{C}} y_{vi} = c_v, y_{vi} \in [0,1], \forall v \in V, i \in \mathcal{C};\ y_{vi} = 1, \forall i \in \mathcal{C}, v \in \mathcal{S}_i\}$, representing the collection of (marginal) probabilities satisfying the capacity and source constraints.

Using the definition of Y, and from the fact that x_{vi}'s are independent and path p is simple (no loop), we observe that

$$D^o(Y,S) = \mathbb{E}_\nu[D^o(X,S)] \ . \quad (9)$$

The extension of D^o to the domain $[0,1]^{|V|\times|C|}$ is known as the multi-linear relaxation of the optimization problem [2], where (6) is relaxed to

$$\min\{D^o(Y,S) : Y \in \mathcal{D}_Y,\, S \in \mathcal{D}_S\} \ . \quad (10)$$

Let X^* and Y^* be the optimal solutions to (6) and (10), respectively. Because the integrality constraints are relaxed, the cost with relaxed variables Y^* satisfies for any $S \in \mathcal{D}_S$:

$$D^o(Y^*,S) \le D^o(X^*,S) \ . \quad (11)$$

The multi-linear relaxation $D^o(Y,S)$ in (9) is non-convex. Therefore, we next approximate it by another cost function D defined as follows:

$$D(Y,S) = \sum_{(i,p)\in\mathcal{R}} \lambda_{(i,p)} D_{(i,p)}(Y,S) \ , \quad (12)$$

where the relaxed delay-cost for request $(i,p) \in \mathcal{R}$ is

$$D_{(i,p)}(Y,S) = \sum_{k=1}^{|p|-1} f(\mathrm{SINR}_{p_{k+1}p_k}(S)) g_{p_k i}(Y) \ , \quad (13)$$

where f is given in (2) and $g_{p_k i}$ is given by

$$g_{p_k i}(Y) = 1 - \min\Big\{1, \sum_{l=1}^{k} y_{p_l i}\Big\}, \quad \forall\, y_{p_l i} \in [0,1]. \quad (14)$$

From the Goemans-Williamson inequality [23], (12) gives an upper bound on (9). Due to the concavity of the $\min$ operator, $\mathbb{E}_\nu\left[g_{p_k i}(Y)\right] \ge g_{p_k i}(\mathbb{E}_\nu[Y])$. In (14), $g_{p_k i}(Y)$ is a piecewise linear function which is not smooth or strictly convex, and its partial derivatives do not exist everywhere. To that end, we devise a subgradient method in Sect. III-D.

The approximated delay-cost $D(Y,S)$ is convex in the caching variables Y due to the convexity of $g_{p_k i}(Y)$. Note that $D(Y,S)$ is nonconvex in S because f is nonconvex in S. We aim to solve the following reduced-complexity formulation (RCF) of the joint optimization problem:

$$\min\{D(Y,S) : Y \in \mathcal{D}_Y,\, S \in \mathcal{D}_S\} \ . \quad (15)$$

The optimal value of $D(Y,S)$ in (15), is within a constant factor from the optimal values of $D^o(Y,S)$ in (10), and of $D^o(X,S)$ in (6). In particular, we have the following theorem.

Theorem 1. Constant factor approximation for fixed S [2], [24]. *For any given $S \in \mathcal{D}_S$, let Y^* and Y^{**} be the optimal solutions that minimize $D^o(Y,S)$ and $D(Y,S)$ in (10) and (15), respectively. Then,*

$$0 \le D^o(Y^{**},S) - D^o(Y^*,S) \le \frac{1}{e}(D^{\mathrm{ub}}(S) - D^o(Y^*,S)) \ .$$

Proof. It follows from relaxing and bounding techniques and

ISBN 978-3-903176-37-9 © 2021 IFIP

by employing Goemans-Williamson inequality [23], [24]. We refer the reader to [2], [5] for existing similar methods. □

b) Rounding: To produce an integral solution to (6), we round the solution Y^{**} of (15). For any given $S \in \mathcal{D}_S$ and given a fractional solution $Y \in \mathcal{D}_Y$, there is always a way to convert it to a $Y' \in \mathcal{D}_Y$ with at least one fewer fractional entry than Y, for which $D^o(Y', S) \leq D^o(Y, S)$ [5]. Each rounding step reduces the number of fractional variables by at least 1. Thus, the above algorithm concludes in at most $|V| \times |C|$ steps (assuming fixed power allocations), producing an integral solution $X' \in \mathcal{D}_X$ such that $D^o(X', S) \leq D^o(Y^{**}, S)$ because each rounding step can only decrease D^o. Hence, from Theorem 1 and (11) we have the following corollary.

Corollary 1. *The integral solution $X' \in \mathcal{D}_X$ as a result of rounding satisfies for any given $S \in \mathcal{D}_S$:*

$$D^o(X^*, S) \leq D^o(X', S) \leq \frac{D^{\rm ub}(S)}{e} + \Big(1 - \frac{1}{e}\Big) D^o(X^*, S) \ .$$

Note that the rounding step produces a $\big(1 - \frac{1}{e}\big)$-approximate solution, along with an offset of $\frac{D^{\rm ub}(S)}{e}$ to RCF. The offset in Cor. 1 is eliminated if instead of RCF in (15) we use a maximum caching gain formulation which concerns the ultimate gain that can be obtained via caching at intermediate nodes, such as in [2] and [5]. In maximizing the caching gain, the objective function is given by the difference $D^{\rm ub}(S) - D(Y, S)$, where $D^{\rm ub}(S)$ is given by (5). However, in this formulation $D^{\rm ub}(S) - D(Y, S)$ increases in S, requiring high powers. Hence, despite its offset, RCF formulation in (15) is preferable as it can jointly optimize power.

c) D^o and D are not jointly convex in Y and S: The transmission delays are coupled due to the interference from simultaneous transmissions. From (2), f is not convex in S. Furthermore, (9) is not convex in Y for given S and not convex in S for given Y, hence not jointly convex in (Y, S). Note that $D(Y, S)$ is jointly convex at low interference or low power because the logarithm function in (2) changes linearly in power when SINR is low in all paths, which is true in the power-limited regime.

For the general setting, the joint convexity of $D(Y, S)$ requires its Hessian matrix H with respect to (Y, S) to be positive semi-definite (PSD). Since (14) is not differentiable, the Hessian matrix for $D(Y, S)$ with respect to Y, i.e., $\nabla^2_Y D$, is not defined. However, from [25, Theorem 2.1], the second order derivatives for maximum functions are defined in each interval and the subhessians of (12) or (14) with respect to Y, i.e., $\{d^2_Y D\}$, exist and we can define a subhessian matrix $d^2_Y D$. However, since (14) is piecewise linear, $d^2_Y D$ is a zero matrix. Combining this with the Schur's complement condition for H to be PSD [26], $D(Y, S)$ is jointly convex only if the off-diagonal blocks of H are singular. However, in our setting, the partial derivatives $\nabla_S D$ with respect to S are nonzero, and the subhessian matrix formed by their subgradients with respect to Y is non-singular. Therefore, in general, $D(Y, S)$ is not jointly convex in (Y, S). Note however that if we define D in the logarithms of the power variables, the function can be biconvex under a certain condition we provide next, in Sect. III-B in Prop. 1.

B. Power Optimization for RCF

We next provide a sufficient condition for $f({\rm SINR}_{p_{k+1}p_k})$ to be convex in $\log$ power variables $P \triangleq (\log(s_{vu}))_{(v,u)\in E}$ in which $P_{vu} = \log(s_{vu})$ denotes power measured on link (v, u) corresponding to request (i, p) in dB.

Proposition 1. Convexity in $\log$ power variables. *A sufficient condition for the composite function $f({\rm SINR}_{p_{k+1}p_k})$ to be convex in $P \triangleq (\log(s_{vu}))_{(v,u)\in E}$ is given as follows.*

$$\frac{2f^{'}(x)^2}{f(x)} \cdot x - f^{'}(x) \leq f^{''}(x) \cdot x, \quad \forall x \geq 0 \ . \qquad (16)$$

Proof. It follows from extending the approach in [27]. □

The sufficient condition (16) of Prop. 1 holds in the high SINR regime where $\log(1 + {\rm SINR}) \approx \log({\rm SINR})$, i.e., where ${\rm SINR} \gg 1$. Given the sufficient condition in (16), it is clear that the program (15) is convex in terms of power measured in dB. Hence, we define the log-power variables P, belonging to the feasible set

$$\mathcal{D}_P = \{P_{vu} \in \mathbb{R} : \sum\nolimits_{u\in O_v} e^{P_{vu}} \leq \hat{s}_v, \ \forall v \in V, \ \forall (u, v) \in E\},$$

where $O_v = \{u \in V : (v, u) \in E\}$.

The condition of Prop. 1 ensures that $D(Y, P)$ is biconvex, i.e., $D(Y, P)$ is convex in Y for given P and convex in P for given Y [28]. However, this condition does not ensure the biconvexity of $D(Y, S)$ because it is nonconvex in S when interference is non-negligible, i.e., at low SINR.

C. General Joint Optimization

We extend the approach of [5] to develop centralized algorithms for the joint power-caching optimization of RCF which is not biconvex, i.e., the sufficient condition in $\log$ powers imposed by Prop. 1 does not hold.

We first present a general result on the relaxed cost function $D(Y, S)$ without putting any assumptions on the $\log$ powers or the caching variables.

a) Strict quasi-convexity of $D(Y, S)$:

Proposition 2. *The relaxed delay-cost function $D(Y, S)$ of RCF in (15) is strictly quasi-convex.*

Proof. Due to space limits we refer the reader to the full version of the work. See [29, Appendix C]. □

The partial derivatives of the relaxed delay-cost function $D_{(i,p)}(Y, S)$, $(i, p) \in \mathcal{R} : (u, v) \in p$ with respect to s_{vu} and the subgradients of $D_{(i,p)}(Y, S)$ with respect to y_{vi} satisfy

$$\frac{\partial D_{(i,p)}}{\partial s_{ju}} \overset{(a)}{\geq} 0, \quad \frac{\partial^2 D_{(i,p)}}{\partial s_{ju}^2} \overset{(a)}{\leq} 0, \ j \in V \backslash v, \qquad (17)$$

$$\frac{\partial D_{(i,p)}}{\partial s_{vu}} \overset{(b)}{\leq} 0, \quad \frac{\partial^2 D_{(i,p)}}{\partial s_{vu}^2} \overset{(b)}{\geq} 0,$$

$$d_{y_{mi}} D_{(i,p)} \overset{(c)}{\leq} 0, \quad d^2_{y_{mi}} D_{(i,p)} \overset{(d)}{\geq} 0, \ m \in p, \qquad (18)$$

where (a) follows from that $f(\mathrm{SINR}_{vu}(S))$ is a decreasing function of $\mathrm{SINR}_{vu}(S)$ which is decreasing in s_{ju} for $j \in V\backslash v$, and similarly (b) from that $\mathrm{SINR}_{vu}(S)$ is linearly proportional to s_{vu} and $f(\mathrm{SINR}_{vu}(S))$ is inversely proportional to $\log(1+\mathrm{SINR}_{vu}(S))$ and convex in $\mathrm{SINR}_{vu}(S)$. Note that (c) is from (14), and (d) from the convexity of $D(Y,S)$ in Y.

The following characterizes the optimality conditions for the relaxed delay-cost function $D(Y,S)$ with a general convex power allocation region $\mathcal{D}_S$ which is true from linearity of (8), and a general convex cache allocation region $\mathcal{D}_Y$.

b) Generalized KKT conditions that requires strictly convex $\mathcal{D}_S$ for unique optimal solution:

Proposition 3. *Assume that the cost functions $D_{(i,p)}(Y,S)$ satisfy (17) and (18), and $\mathcal{D}_S$ is convex. Then, for a feasible set of cache and power allocations $(y_{vi})_{v\in V,\, i\in\mathcal{C}}$ and $(s_{vu})_{(v,u)\in E}$ to be a solution of (15), the following conditions are necessary:*

For all $v \in V$, $i \in \mathcal{C}$, there exists a constant α_{vi} for which

$$\begin{aligned} &d_{y_{vi}}D = \alpha_{vi}, \text{ if } y_{vi} \in (0,1),\\ &d_{y_{vi}}D \geq \alpha_{vi}, \text{ if } y_{vi} = 0,\\ &d_{y_{vi}}D < \alpha_{vi}, \text{ if } y_{vi} = 1, \end{aligned} \tag{19}$$

holds. For all feasible $(\Delta s_{vu})_{(v,u)\in E}$ at $(s_{vu})_{(v,u)\in E}$

$$\sum_{(i,p)\in\mathcal{R}} \frac{\partial D_{(i,p)}}{\partial s_{vu}}(Y,S)\cdot\Delta s_{vu} \geq 0, \tag{20}$$

$$\sum_{(i,p)\in\mathcal{R}} \frac{\partial D_{(i,p)}}{\partial s_{ju}}(Y,S^{**})\cdot\Delta s_{ju} \geq 0,\ j\in V\backslash v, \tag{21}$$

where S^{**} is the optimal power, Δs_{vu} at s_{vu} is an incremental direction that is feasible if there exists $\bar{\delta} > 0$ such that $s_{vu} + \delta\cdot\Delta s_{vu} \in \mathcal{D}_S$, $\forall\ \delta \in (0,\bar{\delta})$.

Proof. See Proof of Prop. 3 in [29]. □

If $D_{(i,p)}(Y,S)$ is jointly convex in (Y,S), the above conditions are also sufficient when (19) holds for all $v \in V$. Furthermore, the optimal S^{**} is unique if $\mathcal{D}_S$ is strictly convex. Moreover, if $D_{(i,p)}(Y,S)$ is strictly convex in Y, then the optimal cache allocations Y^{**} for the relaxed cost function are unique as well. This statement can be proven using arguments similar to the those in [15, Theorem 3].

c) Pareto optimality of $D(Y,S)$: When $f(\mathrm{SINR}_{vu}(S))$ is chosen to be (2), we infer that $D_{(i,p)}(Y,S)$ is in general not jointly convex in (Y,S). Hence, we further need to establish the conditions for a Pareto optimal operating point for strictly quasi-convex cost functions (as shown in Prop. 2). We next show that for a solution (Y^{**}, S^{**}) that both satisfies (19) and (20), we have the following Pareto optimal property.

Theorem 2. Pareto optimality of $D(Y,S)$. *From Prop. 2 on the strict quasi-convexity we have $f(\mathrm{SINR}_{vu}(S))$ in (2), $g_{p_k i}(Y)$ in (14), and the relaxed delay-cost function for RCF in (15) are strictly quasi-convex. If a pair of feasible cache and power allocations $((y^{**}_{vi}),\,(s^{**}_{vu}))$ satisfies conditions (19)-(20) simultaneously, then the vector of transmission delays $(D_{(i,p)}(Y^{**},S^{**}))_{(i,p)\in\mathcal{R}}$ is Pareto optimal, i.e., there does not exist another pair of feasible allocations $((y^{\#}_{vi}),\,(s^{\#}_{vu}))$ such that $D_{(i,p)}(Y^{\#},S^{\#}) \leq D_{(i,p)}(Y^{**},S^{**})$, $\forall (i,p)\in\mathcal{R}$, with at least one inequality being strict.*

Algorithm 1 Projected Subgradient Method

```
Choose $S^0$, $\boldsymbol{y}^0$, small scalar $\epsilon > 0$ and let $t = 0$
do
    Compute subgradient $d_S^t, d_{\boldsymbol{y}}^t$ by (23)
    Determine step sizes $\xi_{\boldsymbol{y}}^t$, $\xi_S^t$ according to (24)
    Compute projected variables $\bar{\boldsymbol{y}}^t$ and $\bar{S}^t$ by (22)
    Update $S^{t+1}$ and $\boldsymbol{y}^{t+1}$ by (22)
    Let $t = t + 1$
while $D^t - D^{t-1} > \epsilon$
Let $(\boldsymbol{y}^*_{sub}, S^*_{sub}) = (\boldsymbol{y}^t, S^t)$
Implement *b) Rounding*.
```

Proof. See [29, Appendix E]. □

Given the relaxed delay-cost function $D(Y,S)$ of the form (12), Theorem 2 implies that at the Pareto optimal point, the cost of a request $(i,p)\in\mathcal{R}$ cannot be strictly reduced without increasing the cost of another request $(i',p')\in\mathcal{R}$.

We next devise a subgradient algorithm to attain the local minima. This is the Pareto optimal solution for $D(Y,S)$ provided that the conditions in Prop. 3 hold. In that case, from Theorem 1, at each rounding step, the subgradients will guarantee a constant factor approximation for any $S \in \mathcal{D}_S$.

D. Joint Caching and Power Optimization via Subgradient

Due to the non-differentiability and non-convexity of $D(Y,S)$ in general SINR condition, we adopt a subgradient projection method solving for the local minima.

a) Algorithm overview: Let $\boldsymbol{y}$ to denote the vectorized caching variable Y, namely $\boldsymbol{y} \in [0,1]^{|V||\mathcal{C}|\times 1}$ with $y_{vi} = \boldsymbol{y}_{(i-1)|V|+v}, \forall v \in V, i \in \mathcal{C}$.

For the t-th iteration, the subgradient projection method can be summarized by the following:

$$\begin{aligned} S^{t+1} &= S^t + \xi_S^t(\bar{S}^t - S^t), \quad \bar{S}^t = [S^t - w_S^t d_S^t]^+_{\mathcal{D}_S},\\ \boldsymbol{y}^{t+1} &= \boldsymbol{y}^t + \xi_{\boldsymbol{y}}^t(\bar{\boldsymbol{y}}^t - \boldsymbol{y}^t), \quad \bar{\boldsymbol{y}}^t = [\boldsymbol{y}^t - w_Y^t d_{\boldsymbol{y}}^t]^+_{\mathcal{D}_{\boldsymbol{y}}}, \end{aligned} \tag{22}$$

where ξ_S^t, $\xi_{\boldsymbol{y}}^t \in (0,1]$ are step sizes respectively corresponding to S and $\boldsymbol{y}$, w_S^t and w_Y^t are positive scalars, $[x]^+_A$ denotes projection of vector x on a convex constraint set A, and

$$d_S^t = \nabla_S D(Y^t,S^t), \quad d_{\boldsymbol{y}}^t \in \partial_{\boldsymbol{y}} D(Y^t,S^t)\ , \tag{23}$$

where d_S^t and $d_{\boldsymbol{y}}^t$ are the subgradients at iteration t with respect to S and $\boldsymbol{y}$, respectively. $\partial_{\boldsymbol{y}} D(Y^t,S^t)$ is the subdifferential with respect to $\boldsymbol{y}$.

b) Subgradient: Note that since $D(Y,S)$ is continuously differentiable in S over set $\mathcal{D}_S$, the subdifferential of $D(Y,S)$ with respect to S will only contain the gradient. Meanwhile, $\partial_{\boldsymbol{y}} D(Y^t,S^t)$ could be explicitly calculated by evaluating $\partial_{y_{vi}} g_{p_k i}$'s inside the term (13) and using (12), where

$$\partial_{y_{vi}} g_{p_k i} = \begin{cases} \{1\}, & \text{if } \sum_{l=1}^k y_{p_l i} < 1\ ,\\ \{0\}, & \text{if } \sum_{l=1}^k y_{p_l i} > 1\ ,\\ [0,1], & \text{if } \sum_{l=1}^k y_{p_l i} = 1\ . \end{cases}$$

ISBN 978-3-903176-37-9 © 2021 IFIP

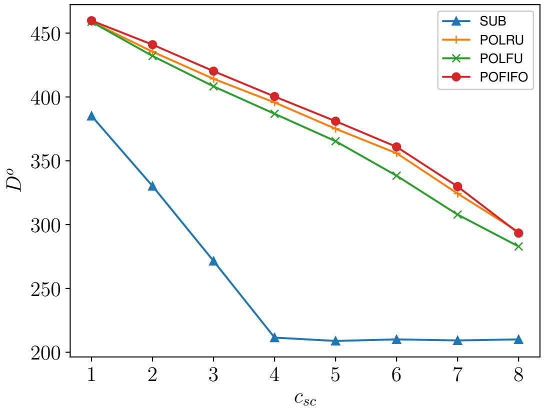

Fig. 2: D^o versus increasing cache capacity. All SC cache capacities are equal to c_{sc}, MC cache capacity is $c_{mc} = \min(2c_{sc}, 8)$, $\gamma = 0.25$ and $\hat{s}_v = 100$.

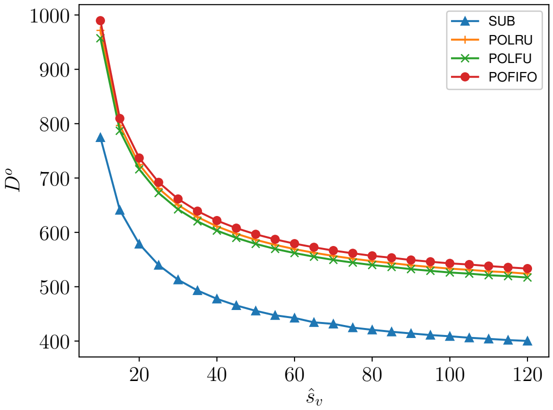

Fig. 3: D^o versus $\hat{s}_v$, which is the same for all $v \in V$. All SC cache capacities are equal to $c_{sc} = 2$, and $c_{mc} = 4$. $\gamma = 0.25$.

c) Step size: The gradient/subgradient magnitudes might be significantly different for Y and S, and thus we compute their step sizes separately. Note that $D(Y, S)$ is not *Lipschitz continuous* in S [30, Sect. 1.2.2], we use a modified Polyak's step size [31]. Let $D^t = D(\boldsymbol{y}^t, S^t)$, then

$$\xi_{\boldsymbol{y}}^t = \frac{D^t - \hat{D}^t}{\|d_{\boldsymbol{y}}^t\|^2}, \quad \xi_S^t = \frac{D^t - \hat{D}^t}{\|d_S^t\|^2} \tag{24}$$

where $\hat{D}^t = \min_{j=0,\cdots,t} D(\boldsymbol{y}^t, S^t) - \delta_t$ is an estimation of the local minima, $\{\delta_t\}_{t\geq 0}$ is a sequence of positive scalars satisfying $\lim_{t\to\infty} \delta_t = 0$, $\lim_{t\to\infty} \sum_{m=0}^{t} \delta_m = \infty$.

d) Convergence: Using the modified Polyak's step size in (24), the subgradient projection algorithm is guaranteed to converge to a local minima D_{sub}^*, which we provide next.

Lemma 1. *Let $(\boldsymbol{y}^t, S^t)$ be generated by the subgradient projection algorithm with modified Polyak's step size (24). Then, the algorithm converges to a local minima D_{sub}^*, i.e.,*

$$\liminf_{t\to\infty} D^t = D_{sub}^* \ . \tag{25}$$

We omit the proof of Lemma 1 as convergence of subgradient projection has been widely studied, e.g., in [32, Ch. 7].

The subgradient projection algorithm converges linearly. To see this, define $(\boldsymbol{y}_{sub}^*, S_{sub}^*)$ to be the set of $\boldsymbol{y}$ and S that attains the local minima D_{sub}^*. Given the objective $D(\boldsymbol{y}, S)$ and its subgradients are bounded near $(\boldsymbol{y}_{sub}^*, S_{sub}^*)$, we say $D(\boldsymbol{y}, S)$ has a *sharp set of minima* near and inside $(\boldsymbol{y}_{sub}^*, S_{sub}^*)$, namely there exists $\mu > 0$ such that for any $S \in \mathcal{D}_S$ and $\boldsymbol{y} \in \mathcal{D}_Y$,

$$D(\boldsymbol{y}, S) - D_{sub}^* \geq \mu L(\boldsymbol{y}, S) \ , \tag{26}$$

where $L(\boldsymbol{y}, S) = \min_{y \in \boldsymbol{y}_{sub}^*,\, s \in S_{sub}^*} \sqrt{\|\boldsymbol{y} - y\|^2 + \|S - s\|^2}$ denotes the distance from $(\boldsymbol{y}, S)$ to $(\boldsymbol{y}_{sub}^*, S_{sub}^*)$.

The existence of *sharp set of minima* further leads to a bound of improvement by each step, i.e.,

$$\left(L^{t+1}\right)^2 \leq \left(L^t\right)^2 - \frac{D^t - D_{sub}^*}{U^2}.$$

where $L^t = L(\boldsymbol{y}^t, S^t)$. Lemma 2 below then follows by aggregating the bounds for iteration 0 through $t - 1$.

Lemma 2. *Let $(\boldsymbol{y}^t, S^t)$ be generated by the subgradient projection algorithm with modified Polyak's step size in (24). Then, L linearly converges according to*

$$L(\boldsymbol{y}^t, S^t) \leq \left(1 - \frac{\mu^2}{U^2}\right)^{\frac{t}{2}} L(\boldsymbol{y}^0, S^0) \ , \tag{27}$$

where μ is the finite positive scalar in (26), and U is a finite positive scalar with $\|d_{\boldsymbol{y}}^t\|^2 + \|d_S^t\|^2 \leq U$ for any t.

Proof. See [29, Appendix G]. □

We summarize the subgradient projection method that achieves the local minima in Algorithm 1.

IV. Numerical Results

In this section, we present numerical results obtained from several simulation scenarios. We simulate a network in accordance with the model in Sect. II and compare the performance of Algorithm 1 (SUB) to the LRU, LFU and FIFO cache replacement policies. We pair these policies with power optimization to have a fair comparison. To make this distinction clear, we name these power optimal (PO) policies POLRU, POLFU and POFIFO when reporting results.

Simulation setup. We simulate a network with 30 users, 4 SCs and a single MC. Users are distributed uniformly, while SCs are distributed using Lloyd's algorithm [33], inside the coverage area of the MC. The users do not cache items, and each one requests a single item at a given time, from a catalog of 10 items, based on a Zipf distribution with parameter γ which can be interpreted as the popularity distribution of content items. The backhaul is the source for all items while the MC and SCs are not designated sources for any item. When a request for an item arrives at the MC or an SC, if the item is not already cached there, it is retrieved from an uplink node that caches the item or from the backhaul and then cached. We calculate gains using pathloss exponent $n = 3.7$ and we set noise power to $N_u = 1$ for all $u \in V$. For SUB, we set the initial points S^0 and Y^0 so that $s_{vu}^0 = \hat{s}_v / |O_v|$ and $y_{vi}^0 = 0$ for all $v, u \in V$ and $i \in \mathcal{C}$. While SUB can optimize a snapshot of the network, LRU, LFU and FIFO policies assume a cache history. Therefore, we simulate these policies in a time-slotted fashion and compare their average results to SUB. We now discuss our observations from three distinct simulation settings. We include any other necessary parameters and details in these discussions.

Effect of cache capacity constraints. We present the results of this setting in Fig. 2. We see that, with increasing cache capacities, our joint optimization algorithm, SUB, reduces delay at a much faster rate compared to traditional replacement policies. SUB also achieves a point of minimum delay given large enough caches, while traditional policies do not converge to such a point and perform worse than SUB with all values of the cache capacity constraint. Numerically, SUB achieves at least 15% less delay, with up to 50% less delay at $c_{sc} = 4$, with the given parameters.

Effect of power constraints. We present the results of this setting in Fig. 3. We observe that traditional policies and SUB show a similar decreasing trend in delay when the total power budget is increased. However, we can still observe the benefit of jointly optimizing power with caching: our algorithm achieves 25% less delay compared to the best performing traditional method, POLFU.

V. Conclusion

We considered the problem of joint power and caching optimization to minimize the transmission delay for a stationary request process in wireless HetNets. Because this problem is NP-complete, we studied several approximation methods that rely on convex relaxation and rounding of caching variables to construct an integral solution. We demonstrated Pareto optimality of the solution to RCF, and devised a subgradient projection algoritm for general non-convex RCF. The results of our approach can enable wireless HetNets to optimally exploit resources to minimize the use of backhaul connections, thus minimizing the transmission delays in both mobile devices and the infrastructure, and to support latency-sensitive applications. They also quantify the potential cost savings from the deployment of SCs. More generally, optimal caching and power control algorithms represent a key enabling technology for realizing the potential of mobile edge computing and fog computing.

References

[1] J. G. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, “What will 5G be?” *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[2] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, “FemtoCaching: Wireless content delivery through distributed caching helpers,” *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–13, Sep. 2013.

[3] H. Che, Y. Tung, and Z. Wang, “Hierarchical web caching systems: Modeling, design and experimental results,” *IEEE J. Sel. Areas Commun.*, vol. 20, no. 7, pp. 1305–14, Sep. 2002.

[4] M. Garetto, E. Leonardi, and V. Martina, “A unified approach to the performance analysis of caching systems,” *ACM Trans. Modeling and Perf. Eval. Comp. Systems*, vol. 1, no. 3, p. 12, May 2016.

[5] S. Ioannidis and E. M. Yeh, “Adaptive caching networks with optimality guarantees,” in *Proc., ACM Sigmetrics*, Jun. 2016, pp. 113 – 124.

[6] M. Mahdian and E. Yeh, “MinDelay: Low-latency joint caching and forwarding for multi-hop networks,” in *Proc., IEEE Int. Conf. Commun. (ICC)*, Kansas City, MO, May 2018, pp. 1–7.

[7] M. A. Maddah-Ali and U. Niesen, “Fundamental limits of caching,” *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856–67, May 2014.

[8] M. Dehghan, A. Seetharam, B. Jiang, T. He, T. Salonidis, J. Kurose, D. Towsley, and R. Sitaraman, “On the complexity of optimal routing and content caching in heterogeneous networks,” in *Proc., IEEE Infocom*, Apr. 2015, pp. 936–944.

[9] N. Abedini and S. Shakkottai, “Content caching and scheduling in wireless networks with elastic and inelastic traffic,” *IEEE/ACM Trans. Netw.*, vol. 22, no. 3, pp. 864–874, May 2013.

[10] Z. Chen, N. Pappas, and M. Kountouris, “Probabilistic caching in wireless D2D networks: Hit optimal vs. throughput optimal,” *IEEE Commun. Letters*, vol. 21, no. 3, pp. 584–587, Mar. 2017.

[11] B. Błaszczyszyn, P. Keeler, and P. Muhlethaler, “Optimizing spatial throughput in device-to-device networks,” in *Proc., IEEE WiOpt*, May 2017.

[12] H. Gupta, N. He, and R. Srikant, “Optimization and learning algorithms for stochastic and adversarial power control,” in *Proc., IEEE WiOpt*, Jun. 2019.

[13] C. Liaskos, X. Dimitropoulos, and L. Tassiulas, “Backpressure on the backbone: A lightweight, non-intrusive traffic engineering approach,” *IEEE Trans. Netw. and Serv. Manag.*, vol. 14, no. 1, pp. 176–190, Nov. 2016.

[14] A. Dvir and N. Carlsson, “Power-aware recovery for geographic routing,” in *Proc., IEEE WCNC*, vol. 5, Apr. 2009, pp. 2851–2856.

[15] Y. Xi and E. M. Yeh, “Node-based optimal power control, routing, and congestion control in wireless networks,” *IEEE Trans. Inf. Theory*, vol. 54, no. 9, pp. 4081–4106, Sep. 2008.

[16] S. Sardellitti, G. Scutari, and S. Barbarossa, “Joint optimization of radio and computational resources for multicell mobile-edge computing,” *IEEE Trans. Signal and Inf. Process. over Netw.*, vol. 1, no. 2, pp. 89–103, Jun. 2015.

[17] J. Liu, Y. Mao, J. Zhang, and K. B. Letaief, “Delay-optimal computation task scheduling for mobile-edge computing systems,” in *Proc., IEEE ISIT*, Jul. 2016, pp. 1451–1455.

[18] R. Gallager, “A minimum delay routing algorithm using distributed computation,” *IEEE Trans. Commun.*, vol. 25, pp. 73–85, Jan. 1977.

[19] J. Oueis, E. C. Strinati, and S. Barbarossa, “The fog balancing: Load distribution for small cell cloud computing,” in *Proc., IEEE VTC*, May 2015.

[20] M. Yemini and A. J. Goldsmith, “Fog optimization via virtual cells in cellular network resource allocation,” *arXiv preprint arXiv:1901.06669*, Jan. 2019.

[21] S. Ioannidis and E. M. Yeh, “Jointly optimal routing and caching for arbitrary network topologies,” in *Proc., ACM ICN*, Sep. 2017.

[22] H. S. Dhillon, R. K. Ganti, F. Baccelli, and J. G. Andrews, “Modeling and analysis of K-tier downlink heterogeneous cellular networks,” *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 550–560, Mar. 2012.

[23] M. X. Goemans and D. P. Williamson, “New 3/4- approximation algorithms for the maximum satisfiability problem,” *SIAM J. Discrete Math.*, vol. 7, no. 4, pp. 656–666, Nov. 1994.

[24] A. A. Ageev and M. I. Sviridenko, “Pipage rounding: A new method of constructing algorithms with proven performance guarantee,” *J. Comb. Optim.*, vol. 8, no. 3, pp. 307–328, Sep. 2004.

[25] S. Scheimberg and P. Oliveira, “Descent algorithm for a class of convex nondifferentiable functions,” *Journal of Optimization Theory and Applications*, vol. 72, no. 2, pp. 269–297, Feb. 1992.

[26] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge university press, 2009.

[27] J. Huang, R. A. Berry, and M. L. Honig, “Distributed interference compensation for wireless networks,” *IEEE J. Sel. Areas Commun.*, vol. 24, no. 5, pp. 1074–1084, May 2006.

[28] C. A. Floudas, *Deterministic Global Optimization: Theory, Methods and Applications*. Springer Science & Business Media, 2013, vol. 37.

[29] D. Malak, F. V. Mutlu, J. Zhang, and E. M. Yeh, “Transmission delay minimization via joint power control and caching in wireless HetNets,” *arxiv preprint arXiv:2105.14380*, May 2021.

[30] D. P. Bertsekas, W. Hager, and O. Mangasarian, *Nonlinear Programming*. Athena Scientific Belmont, MA, 1998.

[31] B. T. Polyak, *Introduction to Optimization*. Optimization Software, Inc., Publications Division, New York, 1987, vol. 1.

[32] S. Boyd, L. Xiao, and A. Mutapcic, “Subgradient methods,” Lecture notes of EE392o, Stanford University, Autumn Quarter, 2003.

[33] S. Lloyd, “Least squares quantization in PCM,” *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.

ISBN 978-3-903176-37-9 © 2021 IFIP