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

Derya Malak, *Member, IEEE*, Faruk Volkan Mutlu, Jinkun Zhang, and Edmund M. Yeh, *Senior Member, IEEE*

***Abstract*— A fundamental challenge in wireless heterogeneous networks (HetNets) is to effectively utilize 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 formulate 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) which are 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 then solve RCF in two ways: 1) alternating optimization of the power and caching variables by leveraging biconvexity, and 2) joint optimization of power control and caching. We characterize the necessary (KKT) conditions for an optimal solution to RCF, and use 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 under general SINR conditions. Finally, our analytical findings are supported by results from extensive numerical experiments.**

***Index Terms*— Wireless joint power-caching optimization, biconvexity, alternating optimization, quasi-convexity, Pareto optimality.**

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.

Derya Malak is with the Communication Systems Department, Ecole d'Ingénieur et Centre de Recherche en Sciences du Numérique (EURECOM), Biot, 06904 Sophia Antipolis, France (e-mail: derya.malak@eurecom.fr).

Faruk Volkan Mutlu, Jinkun Zhang, and Edmund M. Yeh are with the Electrical and Computer Engineering Department, Northeastern University, Boston, MA 02115 USA (e-mail: fvmutlu@ece.neu.edu; zhang.jinku@northeastern.edu; eyeh@ece.neu.edu).

A. Current State of the Art and Motivation

Research to date on cost optimization in the context 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 file downloading 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], in the context of web caches [4], and extension of Che's decoupling approach to provide a unified analysis of caching for different replacement policies in [5]. A simple and ubiquitous algorithm for populating caches in peer-to-peer networking is path replication, i.e., once a request for an item reaches a cache, every downstream node receiving the response caches the item [6]. Various cache eviction policies devised for a single cache primarily concern the optimization of the cache hit rate that describes the frequency of finding the searched item in the cache, or the latency that describes how long it takes for the cache to return a desired item [3], [5], [7].

For networks of caches, time-to-live (TTL) caching is a better alternative [5], [8], where items stay in a cache for predetermined times and are evicted when the timers expire. An age-based-threshold policy where cache stores all contents requested more times than a threshold [9] captures temporal popularity changes via the Poisson shot noise model (SNM), and maximizes the hit ratio [10]. Hence, SNM is compatible with the TTL caching [11]. Traditional cache eviction policies [3], [5], [12], e.g., LRU, Least Frequently Used (LFU), First-In-First-Out (FIFO), and Random Replacement (RR), provide gain by making content available locally and compromise between hit rate and latency, and can be arbitrarily suboptimal in terms of the expected caching gain [13]. However, as devised in the landmark paper [14], novel coded caching

approaches can provide a global gain that derives from jointly optimizing placement and delivery. Furthermore, geographic caching approaches that capture the spatial diversity of content, as in [15], [16], [17], and [18], help optimize the placement.

There is an extensive literature on physical layer aspects of caching in wireless networks design. For example, the gain offered by local caching and broadcasting is characterized in [14]. Works also include the analysis of the scaling of the per-user throughput and collaboration distance [17], the wireless caching capacity region which is the closure of the set of all achievable caching traffic [19], as well as single-hop and device-to-device [14], [17], [20], [21], [22], and multi-hop caching networks [23], [24].

Recently, information centric networking (ICN) architectures have put emphasis on the traffic engineering and caching problems [13], [25] to effectively use both bandwidth and storage for efficient content distribution [26], and optimize the network performance [27]. Alternatively, there have been works focusing on jointly optimizing the caching gain and resource usage, e.g., a decentralized SC caching optimization, i.e., femtocaching, to minimize the download delay [2], distributed optimization of caching gain given routing [13], minimizing the total cost incurred in storing and accessing objects by building the Steiner trees [28], jointly optimizing caching and routing to provide latency guarantees [29], and minimizing delay by taking into account congestion [30], and elastic and inelastic traffic [31]. Existing strategies have also focused on separately optimizing the caching gain or the throughput [32], or spatial throughput via scheduling [33]. 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 [34], [35], [36], [37], [38], or maximize throughput under latency considerations [39]. 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 [40] and [41], joint optimization of power control, routing, and congestion [42], and joint optimization of radio and computational resources under latency and power constraints [43], [44], [45], as well as delay-optimal computation task scheduling at the mobile edge [46], and the minimum delay routing algorithm [47]. In addition, fog optimization-based resource allocation schemes for wireless networks have been devised in [48] to achieve high power efficiency while keeping a very high Quality of Experience under latency constraints, and in [49] to maximize the sum rate of cellular networks. However, none of these or research on ICN architectures has jointly designed traffic engineering and cache placement strategies to optimize network performance in view of traffic demands.

Several papers have studied complexity and optimization issues of cost minimization as an offline, centralized caching problem under restricted topologies [2], [6], [28], [30], [50], [51]. Despite the advent of different caching solutions, to the best of our knowledge, none of the above protocols focuses on the joint optimization of caching and power allocation or provides algorithmic performance guarantees in terms of the achievable costs via caching. Although most of these strategies suggest that intermediate caching can alleviate the average download delays, 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.[1]

B. Methodology and Contributions

In this paper, we study jointly optimal caching and power control for arbitrary multi-hop wireless HetNet topologies with nodes that have caching capabilities. Note that as the networks are becoming increasingly heterogeneous, MCs and SCs can co-exist in 5G, and all networks beyond it [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,[2] a set of finite cache storage capacities, a demand distribution on the content items known a priori, and a subset of nodes designated to store specific items, we devise algorithms 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 delay, propagation delay, processing delay and queuing delay, we are primarily interested in a lightly loaded regime for which congestion-dependent latency costs can be neglected, and in which the link lengths are much smaller than the propagation speed of the signal, and each node can sustain a high service rate relative to the average rate at which items are arriving to be serviced. Hence the transmission delay is the major delay component. 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. We emphasize that 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], [13], [24], and [52] 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,

[1] In this paper, we primarily consider the transmission delay assuming a lightly loaded system which we detail in Sect. II.

[2] Routing is fixed and each request is a pair that is jointly determined by the item requested and the fixed multi-hop path traversed to serve this request.

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. We jointly optimize RCF under the biconvexity condition to provide an alternating optimization solution to minimizing $D(Y,S)$.
- **Joint optimization framework.** We jointly optimize RCF under the general setting which is not jointly convex. We obtain the following results: **a)** quasi-convexity of $D(Y,S)$, **b)** necessary conditions for optimality of $D(Y,S)$, **c)** generalized necessary conditions for optimality of $D(Y,S)$ assuming strict convexity of the feasible set of all S, denoted by $\mathcal{D}_S$, and **d)** Pareto optimality of the solution to $D(Y,S)$.
- **Subgradient projection algorithm.** Due to the non-differentiability and non-convexity of the relaxed problem, we propose a Clarke subgradient projection algorithm with a modified Polyak's step size, and provide a simple method to calculate the Clarke subgradients.

Our prior work [53] contains a subset of the results of this manuscript, including (i) a RCF $D(Y,S)$ to the joint power-caching optimization problem, (ii) sufficient conditions for the biconvexity of $D(Y,S)$, (iii) a joint optimization framework, where we show a) 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 $D(Y,S)$, and (iv) a subgradient projection algorithm. However, [53] does not contain the results pertaining to the alternating optimization. It does not include the proofs of the main results (theorems, or propositions), which have been detailed in the current manuscript. This draft contains in addition to a)-c) under the joint optimization framework, necessary conditions for the optimality of $D(Y,S)$ (Theorem 2). The simulation results in the current manuscript are also more comprehensive and applicable to larger scale models.

Organization of the rest of the paper is follows. In Sect. II we detail the wireless HetNet topology where each node has caching capability and adjustable transmission powers. We establish a transmission delay model of serving a request using multiple hops where the transmission delay is a nonlinear function of the signal-to-interference-plus-noise ratios (SINR). In Sect. III we detail the joint optimization of delay in power and caching variables. This section contains the main technical contributions which are the necessary and sufficient conditions for joint optimality, and algorithms to solve for points with provable theoretical guarantees. In Sect. IV, we numerically verify our analytical findings. In Sect. V, we conclude the paper by pointing out the use cases including mobile edge and fog computing.

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. To alleviate the impact of limited backhaul capacity, availability, and long-distance reach it is desired that the network serves the requests via the SCs and multi-hop transmissions. While each MC or SC might have a fiber connection to the backhaul network in 5G, multi-hop relaying[3] is essential due to radio range limitations. However, for a given end-to-end distance, increasing the number of hops arbitrarily may lead to additional energy consumption incurred by relays. As a result, long-hop routing, sending over a smaller number of longer hops versus over many short hops, is a competitive strategy for many networks [56]. Furthermore, from a cost-effective perspective, each MC should allocate its resources to a smaller number of users, which balances the traffic between SCs and MCs [1]. We represent the network as a directed graph $\mathcal{G}(V,E)$ where V is the collection 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,[4] 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 proposed wireless caching network model and possible multi-hop paths where the users request different items. We provide the notation in Table I.

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. Readers can refer to [58, Section 6] for an extension to contents with unequal sizes where the authors partitioned contents into equal-sized chunks and defined the caching gain/cost per chunk, where chunks can be treated as distinct items [52]. 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: $x_{vi} = 1,\ \forall v \in \mathcal{S}_i$. While the backhaul is always considered to be a designated source for all items, user nodes, SCs, or MCs can also be designated

[3]Since the transceiver is the major source of power consumption in a node and long distance transmission requires high power, in some cases multi-hop routing can be more energy efficient than single-hop routing [54], [55].

[4]If subsets of nodes are allocated different frequencies, as in OFDMA-based networks, then we can determine the resulting subset of interfering nodes [57]. This also reduces interference and improves the SINR coverage.

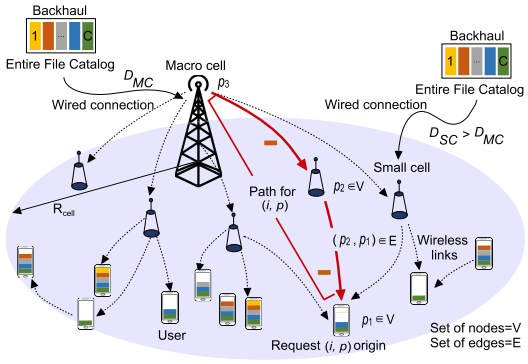


Fig. 1. A caching network scenario with possible connections between the users, SCs or MCs, and to the backhaul, where the backhaul cost D_{SC} of SC is typically higher than the backhaul cost D_{MC} of MC connections [39]. A path $p = \{p_1, p_2, p_3\}$ for request (i, p) is indicated where p_1 is a user where the request (i, p) is originated, p_2 is an SC, and p_3 is the MC.

sources. Items that are not available in the wireless network need to be retrieved from the backhaul via the 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) that is jointly determined by the item $i \in \mathcal{C}$ being requested, and the fixed path p traversed (request is forwarded from the user toward a designated source over a fixed path) to serve this request. The routing strategy of a user with respect to request $(i, p) \in \mathcal{R}$ 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 to a request (i, p) travels the same path p, in the reverse direction, which follows from the symmetric routing assumption in ICN,[5] (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. This is due to the assumption that sizes of requests are much smaller than that of responses, and we ignore the processing time at the nodes.

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 focus on lightly loaded systems, where the transmission delay is dominant and the other delay components are negligible. We assume there is one queue for each link $(u, v) \in E$ that serves in a FIFO manner all requests traversing (u, v).

To determine the transmission delay of link $(v, u) \in E$ corresponding to request (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. Hence, 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 $n > 2$ [59]. The signal for request (i, p) over link (v, u) is decoded regarding all other signals as noise, for all $(i, p) \in \mathcal{R}$ and $(v, u) \in E$. Thus, in our model the transmission delays are coupled, in contrast to [2] and [24], 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 active 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.[6] However, we leave this extension to future work.

To model the wireless transmission delay on link $(v, u) \in E$, we use the following composite relation[7]:

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

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 transmission 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.

If the last node of a path p, i.e., $p_{|p|}$, is not a designated user node, SC, or MC, then the final hop from $p_{|p|-1}$ to $p_{|p|}$ is a wired connection from an MC or an SC to the backhaul. For given $(i, p) \in \mathcal{R}$, the transmission delay incurred by the edge $(p_{|p|-1}, p_{|p|}) \in p$ where $p_{|p|}$ is the backhaul is given as

$$f(\mathrm{SINR}_{p_{|p|} p_{|p|-1}}(S)) = \begin{cases} D_{MC}, & \text{if } p_{|p|-1} \text{ is MC}, \\ D_{SC}, & \text{if } p_{|p|-1} \text{ is SC}, \end{cases} \quad (3)$$

[5] In ICN, since there is no source or destination information in packets, responses follow the reverse path of requests [13].

[6] In 5G and 6G applications, the performance can be improved using interference cancellation techniques with NOMA-based non-orthogonal resource allocation. However, our transmission delay (2) is related to the inverse of transmission rate. This paper focus on the delay tradeoffs due to the joint optimization of power and caching variables. A similar tradeoff for NOMA-based frameworks will be investigated in our future work.

[7] Practical adaptive modulation and coding schemes operate at lower SINR values [60, Ch. 4.2, Ch. 9.3]. For example, for MQAM the gap from the Shannon SNR as function of the symbol error probability P_e is $\Gamma = \frac{1}{3}(Q^{-1}(P_e))^2$.

where the wired backhaul transmission delays D_{MC} and D_{SC} are fixed and known a priori, which we assume to be the same for all SCs and MCs based on [61]. In contradistinction to this, the transmission delays are coupled in the wireless part of the network due to the dependency of $\mathrm{SINR}_{vu}(S)$, $(v,u) \neq (p_{|p|}, p_{|p|-1})$ in (1) on the power allocation 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

In this section, 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. Later in Sect. III-C we jointly optimize RCF, first under the assumption of biconvexity so as to provide an alternating optimization formulation, and second under the general setting which is not jointly convex, we provide various results on the RCF objective. We demonstrate **a)** quasi-convexity of $D(Y,S)$, **b)** necessary conditions for optimality of $D(Y,S)$, **c)** generalized necessary conditions for optimality of $D(Y,S)$ under strict convexity of $\mathcal{D}_S$, and **d)** Pareto optimality of the solution to $D(Y,S)$. Finally in Sect. III-D we provide a subgradient projection algorithm for the general setting.

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

$$D^o_{(i,p)}(X,S) = \sum_{k=1}^{|p|-1} f(\mathrm{SINR}_{p_{k+1}p_k}(S)) \prod_{l=1}^{k}(1-x_{p_l i}) \quad (4)$$

where $x_{vi}=1$ if node $v \in V$ stores item $i \in \mathcal{C}$, and $x_{vi}=0$ otherwise, and $D^o_{(i,p)}(X,S)$ 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. If the request is well-routed, no edge (or cache) appears twice in (4). 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

$$D^o(X,S) = \sum_{(i,p)\in\mathcal{R}} \lambda_{(i,p)} D^o_{(i,p)}(X,S). \quad (5)$$

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.,

$$D^{\mathrm{ub}}(S) = \sum_{(i,p)\in\mathcal{R}} \lambda_{(i,p)} \sum_{k=1}^{|p|-1} f(\mathrm{SINR}_{p_{k+1}p_k}(S)). \quad (6)$$

Our primary objective is to solve the problem

$$\min\{D^o(X,S) : X \in \mathcal{D}_X,\ S \in \mathcal{D}_S\}, \quad (7)$$

where $\mathcal{D}_X$ is the feasible set of $X \in \mathbb{R}^{|V|\times|\mathcal{C}|}$ satisfying the capacity, integrality, and source constraints:

$$\mathcal{D}_X = \Big\{ \sum_{i\in\mathcal{C}} x_{vi} \le c_v,\ \forall v \in V,\ x_{vi} \in \{0,1\},\ v \in V,\ i \in \mathcal{C};$$
$$x_{vi} = 1,\ \forall i \in \mathcal{C},\ v \in \mathcal{S}_i \Big\}. \quad (8)$$

The set of constraints $\mathcal{D}_S$ is for the power or resource budget. The feasible set of S is specified by the individual power budget for each node, namely $\mathcal{D}_S$ is the feasible set of all $S = [s_{vu}]_{v\in V, u\in V\setminus v} \in \mathbb{R}^{|V|\times(|V|-1)}$ satisfying

$$\mathcal{D}_S = \Big\{ \sum_{u\in O_v} s_{vu} \le \hat{s}_v,\ s_{vu} \ge 0,\ \forall v \in V \Big\}, \quad (9)$$

where $O_v = \{u \in V : (v,u) \in E\}$, and $\hat{s}_v$ denotes the total transmit power of node $v \in V$.

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

1) Convex Relaxation: To approximate the non-convex function $D^o(X,S)$, we construct a convex relaxation, following the approach of [2] and [13]. 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 $\{0,1\}^{|V|\times|\mathcal{C}|}$, 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 in (8), let marginal probabilities

$$y_{vi} = \mathbb{P}_\nu[x_{vi}=1] = \mathbb{E}_\nu[x_{vi}] \in [0,1], \quad v \in V, \quad i \in \mathcal{C}. \quad (10)$$

Denote the feasible set of $Y = [y_{vi}]_{v\in V, i\in\mathcal{C}} \in \mathbb{R}^{|V|\times|\mathcal{C}|}$ by

$$\mathcal{D}_Y = \Big\{ \sum_{i\in\mathcal{C}} y_{vi} \le c_v,\ v \in V,\ y_{vi} \in [0,1],\ v \in V,\ i \in \mathcal{C};$$
$$y_{vi} = 1,\ v \in \mathcal{S}_i,\ i \in \mathcal{C} \Big\}, \quad (11)$$

representing the collection of (marginal) probabilities that $v \in V$ stores $i \in \mathcal{C}$ and satisfying the capacity and source constraints for the caching variables. Using the definition of Y in (10), and from the fact that x_{vi}'s are independent and path p is simple (no loop), we now observe that

$$D^o(Y,S) = \mathbb{E}_\nu[D^o(X,S)]. \quad (12)$$

The extension of D^o to the domain $[0,1]^{|V|\times|\mathcal{C}|}$ is known as the multi-linear relaxation of the optimization problem [2], where (7) is relaxed to

$$\min\{D^o(Y,S) : Y \in \mathcal{D}_Y,\ S \in \mathcal{D}_S\}. \quad (13)$$

TABLE I
NOTATION

Definition	Symbol
Cache capacity of $v \in V$; Catalog size; Binary caching variables	c_v ; $\lvert\mathcal{C}\rvert$; $X = [x_{vi}]_{v\in V,\, i\in\mathcal{C}}$
Path of length $\lvert p\rvert = K$ corresponding to request $r = (i, p)$	$p = \{p_1, \ldots, p_K\}$, $p_k \in V$
Arrival rate of request $(i, p) \in \mathcal{R}$; Requests of different types (item, path)	$\lambda_{(i,p)} > 0$; $(i, p) \in \mathcal{R}$
Distance from node v to node u; Path loss exponent	r_{vu}; $n > 2$
Designated sources for $i \in \mathcal{C}$	$\mathcal{S}_i \subseteq V$
Total transmit power of node $v \in V$; Noise power at receiver $u \in V$	$\hat{s}_v$; N_u
SINR function on link $(v, u) \in E$; Delay function on link $(v, u) \in E$	$\mathrm{SINR}_{vu}(S)$; $f(\mathrm{SINR}_{vu}(S)) \geq 0$
Global minimum objective and solution of the original problem (7)	D^*; (Y^*, S^*)
Global minimum objective and solution of the RCF problem (18)	D^{**}; (Y^{**}, S^{**})
Objective and solution of (18) generated by Algorithm 2	D^*_{sub}; $(\boldsymbol{y}^*_{sub}, S^*_{sub})$

Let X^* and Y^* be the optimal solutions to (7) and (13), respectively. Then, because the integrality constraints are relaxed in (10), the cost with relaxed variables Y^* satisfies for any $S \in \mathcal{D}_S$:

$$D^o(Y^*, S) \leq D^o(X^*, S). \tag{14}$$

Note that the multi-linear relaxation $D^o(Y, S)$ in (12) is non-convex. Therefore, we next approximate it by another cost function D defined as follows:

$$D(Y, S) = \sum_{(i,p)\in\mathcal{R}} \lambda_{(i,p)} D_{(i,p)}(Y, S), \tag{15}$$

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

$$D_{(i,p)}(Y, S) = \sum_{k=1}^{|p|-1} f(\mathrm{SINR}_{p_{k+1}p_k}(S)) g_{p_k i}(Y), \tag{16}$$

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

$$g_{p_k i}(Y) = 1 - \min\Big\{1, \sum_{l=1}^{k} y_{p_l i}\Big\}, \quad \forall\, y_{p_l i} \in [0, 1]. \tag{17}$$

The Goemans-Williamson inequality [62] states that,

$$1 - \prod_{l=1}^{k}(1 - y_{p_l i}) \geq \min\{1, \sum_{l=1}^{k} y_{p_l i}\} \quad y_{p_l i} \in [0, 1].$$

Combining (17) with the above, it holds that

$$g_{p_k i}(Y) \geq \prod_{l=1}^{k}(1 - \mathbb{E}_\nu[x_{p_l i}]).$$

Therefore, the relaxed objective (15) gives an upper bound on (12). Due to the concavity of the min operator, i.e., $\mathbb{E}_\nu\left[g_{p_k i}(Y)\right] \geq g_{p_k i}(\mathbb{E}_\nu[Y])$, the function $g_{p_k i}(Y)$ is quasi-convex (see Prop. 2) in Y. In (17), $g_{p_k i}(Y)$ is a piecewise linear function which is not smooth or strictly convex, and its partial derivatives[8] do not exist everywhere. If the objective function or some of the constraint functions are non-differentiable, we can devise non-differentiable methods to optimize $D(Y, S)$, or subdifferential versions of KKT conditions [64], [65, Ch. 6.3]. To address such scenarios we will detail an algorithm in Sect. III-D.

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

$$\min\{D(Y, S) : Y \in \mathcal{D}_Y,\, S \in \mathcal{D}_S\} \tag{18}$$

where the objective function in (18) captures the backhaul connection between the core network and the MCs or the SCs at the edge of the network. This is enabled by incorporating the backhaul delay model in (3) which affects the delay-cost function $D(Y, S)$ given in (15).

The optimal value of $D(Y, S)$ in (18), is guaranteed to be within a constant factor from the optimal values of $D^o(Y, S)$ in (13), and of $D^o(X, S)$ in (7). In particular, we have the following theorem.

Theorem 1: **Constant factor approximation for fixed** S **[2], [66].** *For given S, let Y^* and Y^{**} be the optimal solutions that minimize $D^o(Y, S)$ and $D(Y, S)$ in (13) and (18), respectively. Then,*

$$D^o(Y^*, S) \leq D^o(Y^{**}, S) \leq \frac{D^{\mathrm{ub}}(S)}{e} + \Big(1 - \frac{1}{e}\Big) D^o(Y^*, S). \tag{19}$$

Proof: See Appendix A. □

2) Rounding: To produce an integral solution to (7), we round the solution Y^{**} of (18). 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 $D^o(Y', S) \leq D^o(Y, S)$ [13], [15].

Each rounding step reduces the number of fractional variables by at least 1. Thus, the above algorithm concludes in at most $|V| \times |\mathcal{C}|$ 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 (14) we have the following corollary.

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

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

Note that the rounding step produces a $\big(1 - \frac{1}{e}\big)$-approximate solution, along with an offset of $\frac{D^{\mathrm{ub}}(S)}{e}$ to RCF. The offset in Cor. 1 is eliminated if instead of RCF in (18) 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 [13]. In maximizing the caching gain, the objective function is given by the differ-

[8] A function is piecewise continuously differentiable if each piece is differentiable throughout its subdomain, even if the whole function may not be differentiable at the points between the pieces [63, Ch. 3].

ence $D^{\text{ub}}(S) - D(Y,S)$, where $D^{\text{ub}}(S)$ is given by (6). In this case, the relationship $D^o(X^*,S) \leq D^o(Y^{**},S) \leq \frac{D^{\text{ub}}(S)}{e} + \left(1-\frac{1}{e}\right)D^o(X^*,S)$ is equivalent to $D^{\text{ub}}(S) - D^o(X^*,S) \geq D^{\text{ub}}(S) - D^o(Y^{**},S) \geq D^{\text{ub}}(S) - \frac{D^{\text{ub}}(S)}{e} - \left(1-\frac{1}{e}\right)D^o(X^*,S) = \left(1-\frac{1}{e}\right)(D^{\text{ub}}(S) - D^o(X^*,S))$, giving a $\left(1-\frac{1}{e}\right)$-approximate solution for the maximum caching gain formulation without an offset. However, in this formulation the gap $D^{\text{ub}}(S) - D(Y,S)$ that models the caching gain increases in S, requiring high powers. Hence, despite its offset, RCF in (18) is preferable as it can jointly optimize power.

3) 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, (12) 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 (and its reciprocal is convex) in power when SINR is low in all paths, which is true in the power-limited regime.

The joint convexity of D requires the Hessian matrix H of $D(Y,S)$ with respect to (Y,S) to be positive semi-definite (PSD). Since (17) 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 [67, Theorem 2.1], the second order derivatives for maximum functions are defined in each interval and the subhessians of (15) or (17) with respect to Y, i.e., $\{d^2_Y D\}$, exist and we can define a subhessian matrix $d^2_Y D$. However, since (17) is piecewise linear, $d^2_Y D$ is a zero matrix. Combining this with the Schur's complement condition for H to be PSD in [68], $D(Y,S)$ is jointly convex only if the off-diagonals 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, $D(Y,S)$ is not jointly convex.

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 $f(\text{SINR}_{p_{k+1}p_k})$ 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 $f(\text{SINR}_{p_{k+1}p_k})$ to be convex in $P \triangleq (\log(s_{vu}))_{(v,u)\in E}$ is given as follows.*

$$\frac{2f'(x)^2}{f(x)} \cdot x - f'(x) \leq f''(x) \cdot x, \quad \forall x \geq 0. \tag{20}$$

Proof: The result follows from extending the approach in [69]. For details, see Appendix B. □

The sufficient condition (20) of Prop. 1 holds in the high SINR regime where $\log(1+\text{SINR}) \approx \log(\text{SINR})$, i.e., where $\text{SINR} \gg 1$. Given the sufficient condition in (20), it is clear that the program (18) is convex in terms of power measured in dB. Hence, we define the log-power variables P, belonging to the feasible set

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

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 [70]. This paves the way for employing methods to solve RCF in (18). We next outline one such method.

C. Joint Optimization of RCF

In this section, we present two techniques to optimize RCF: **1)** Biconvex optimization of $D(Y,S)$ under the condition of Prop. 1 on convexity in log powers, and **2)** General joint optimization where $D(Y,S)$ is not jointly convex in Y and S. For the former, we exploit alternating optimization methods. For the latter, we prove various results on $D(Y,S)$: **a)** quasi-convexity, **b)** necessary conditions for optimality, **c)** generalized necessary conditions under strict convexity of $\mathcal{D}_S$, and **d)** Pareto optimality of $D(Y,S)$.

1) Alternating Optimization: From [71, Theorem 4.2], since D is a differentiable and biconvex function of (Y,S), each stationary point of D is a partial optimum. Furthermore, from [71, Corollary 4.3], (Y,S) is stationary if and only if it is a partial optimum. However, a partial optimum neither has to be a global nor a local optimum to the given biconvex optimization problem even if (Y,S) is stationary, as stationary points can be saddle points of D [71]. From (21) and (22) the partial derivatives of $D_{(i,p)}$ does not change sign. However, since the partial derivative of D is a linear combination of $D_{(i,p)}$ as given in (15), it is possible that the stationary point may be a local optimum.

We next present a biconvex optimization technique for RCF. To that end, we exploit alternating optimization methods. There exist techniques to find the local optimum of biconvex minimization problems, such as block-relaxation methods [71]. Furthermore, the global optimum of biconvex problems can be determined for certain classes of constraints [72].

Provided that the convexity condition in Prop. 1 holds, $D(Y,S)$ is biconvex and hence we can focus on the alternating optimization of RCF. This corresponds to alternatively updating the power variables S given the caching variables Y, and then updating Y given S. This iterative optimization approach can find a local optimum to the average delay minimization problem. To obtain an integral solution, the algorithm needs a rounding step before it terminates. This technique for RCF is summarized in Algorithm 1. An algorithm called Global OPtimization (GOP) algorithm was developed in [72] to exploit the convex substructure of constrained biconvex minimization problems by a primal-relaxed dual approach. The objective function and the constraints in RCF satisfy the necessary convexity conditions [70, Ch. 3.1, Conditions (A)] for the GOP algorithm. However, [70, Ch. 3.1], [72, Theorem 1, Condition (d)] require the multipliers for the primal problem to be uniformly bounded, which may not be true for RCF. Hence, employing the GOP algorithm does not guarantee termination in a finite number of steps for any $\epsilon > 0$ [70, Theorem 3.6.1], or at the global optimum of (15) [70, Theorem 3.6.2].

Algorithm 1 Alternating Optimization for Biconvex $D(Y,S)$

Begin: $S^0 \in \mathcal{D}_S; Y^0 \in \mathcal{D}_Y$**;**
Let $t=0$;
do
  $Y^{t+1} = \arg\min_{Y} D(Y, S^t)$ (convex with start point Y^t);
  $S^{t+1} = \arg\min_{S} D(Y^{t+1}, S)$ (convex with start point S^t);
  Let $t = t+1$;
while $D(Y^t, S^t) - D(Y^{t-1}, S^{t-1}) > \epsilon$
Let $(Y^{**}, S^{**}) = (Y^t, S^t)$;
Implement *b) Rounding*.

We note that the proposed alternating approach requires the condition in Prop. 1, while no optimality guarantee is established. 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. Deriving the necessary conditions for optimality will reveal the true potential of the algorithm and elucidate the effect of network's operating regime, e.g., in the high or low SINR.

2) General Joint Optimization: We next extend the approach of [13] 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) Quasi-convexity of D(Y, S):

Proposition 2: The relaxed delay-cost function $D(Y,S)$ of RCF in (18) is quasi-convex.

Proof: See Appendix C. □

Note that 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 it with respect to y_{vi} satisfy

$$\frac{\partial D_{(i,p)}}{\partial s_{ju}} \overset{(a)}{\geq} 0, \quad \frac{\partial^2 D_{(i,p)}}{\partial s_{ju}^2} \overset{(a)}{\leq} 0, \ j \in V \backslash v, \tag{21}$$
$$\frac{\partial D_{(i,p)}}{\partial s_{vu}} \overset{(b)}{\leq} 0, \quad \frac{\partial^2 D_{(i,p)}}{\partial s_{vu}^2} \overset{(b)}{\geq} 0,$$
$$d_{y_{mi}} D_{(i,p)} \overset{(c)}{\leq} 0, \quad d^2_{y_{mi}} D_{(i,p)} \overset{(d)}{\geq} 0, \ m \in p, \tag{22}$$

where (a) follows from that $f(\text{SINR}_{vu}(S))$ is decreasing in $\text{SINR}_{vu}(S)$ which is decreasing in s_{ju} for $j \in V\backslash v$, and similarly (b) from that $\text{SINR}_{vu}(S)$ is linearly proportional to s_{vu} and $f(\text{SINR}_{vu}(S))$ is inversely proportional to $\log(1+\text{SINR}_{vu}(S))$ and convex in $\text{SINR}_{vu}(S)$. Note that (c) follows from (17), and (d) from the convexity of $D(Y,S)$ in Y.

b) Necessary conditions for optimality of D(Y, S): We investigate the necessary, i.e., the Karush-Kuhn-Tucker (KKT), conditions for a solution of $D(Y,S)$ to be optimal. Assume that $D(Y,S)$ and the constraints are continuously differentiable at (Y^{**}, S^{**}). If (Y^{**}, S^{**}) gives a local optimum and the optimization problem satisfies some regularity conditions [65], then there exist constants $[\mu_{v,i}]_{v\in V, i\in\mathcal{C}}$, $[\nu_{v,i}]_{v\in V, i\in\mathcal{C}}$, $[\eta_v]_{v\in V}$, $[\beta_v]_{v\in V}$, $[\gamma_{e,r}]_{e=(u,v)\in E, r=(i,p)\in\mathcal{R}}$ called KKT multipliers, such that the following hold.

Theorem 2: **Necessary conditions for optimality of** $D(Y,S)$**.** *For a feasible set of power and cache allocations $[s_{vu}]_{(u,v)\in E}$, and $[y_{vi}]_{v\in V, i\in\mathcal{C}}$ to be the solution of RCF in (18), the following conditions are necessary.*

The subgradients for the caching variables satisfy

$$\begin{aligned} &d_{y_{vi}} D = \alpha_{vi}, \text{ if } y_{vi} \in (0,1), \\ &d_{y_{vi}} D \geq \alpha_{vi}, \text{ if } y_{vi} = 0, \\ &d_{y_{vi}} D < \alpha_{vi}, \text{ if } y_{vi} = 1, \end{aligned} \tag{23}$$

where α_{vi}, $v \in V$, $i \in \mathcal{C}$ is some constant.

The gradients for the power variables should satisfy

$$\begin{aligned} &\frac{\partial D}{\partial s_{vu}} \geq -\beta_v + \gamma_{e,r}, \text{ if } s_{vu} = 0, \\ &\frac{\partial D}{\partial s_{vu}} = -\beta_v, \text{ if } s_{vu} > 0 \text{ and } \sum_{u\in O_v} s_{vu} = \hat{s}_v, \\ &\frac{\partial D}{\partial s_{vu}} = 0, \text{ if } s_{vu} > 0 \text{ and } \sum_{u\in O_v} s_{vu} < \hat{s}_v, \end{aligned} \tag{24}$$

for nonnegative constants β_v, $v \in V$, $\gamma_{e,r}$, $e=(u,v), r \in \mathcal{R}$.

Proof: See Appendix D. □

Note that when $D_{(i,p)}(Y,S)$ in (16) is jointly convex in (Y,S) (which is not true in general and requires a more restrictive condition than Prop. 1 on the power control variables), the conditions in Theorem 2 are also sufficient for optimality of $D(Y,S)$ [42, Theorem 1].

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 (9), and a general convex cache allocation region $\mathcal{D}_Y$.

c) 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 (21) and (22), 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 (18), the following conditions are necessary:

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

$$\begin{aligned} &d_{y_{vi}} D = \alpha_{vi}, \text{ if } y_{vi} \in (0,1), \\ &d_{y_{vi}} D \geq \alpha_{vi}, \text{ if } y_{vi} = 0, \\ &d_{y_{vi}} D < \alpha_{vi}, \text{ if } y_{vi} = 1. \end{aligned} \tag{25}$$

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

$$\sum_{(i,p)\in\mathcal{R}} \frac{\partial D_{(i,p)}}{\partial s_{vu}}(Y,S) \cdot \Delta s_{vu} \geq 0, \tag{26}$$

$$\sum_{(i,p)\in\mathcal{R}} \frac{\partial D_{(i,p)}}{\partial s_{ju}}(Y,S^{**}) \cdot \Delta s_{ju} \geq 0, \ j \in V\backslash v, \tag{27}$$

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

Proof: The necessary conditions follow from the arguments in Theorem 2. However, we still need to detail why (26) is true. By the convexity of cost functions, the cost difference of two configurations (Y, S^a) and (Y, S^{**}) for

any feasible S^a is

$$\sum_{(i,p)\in\mathcal{R}} D_{(i,p)}(Y,S^a) - \sum_{(i,p)\in\mathcal{R}} D_{(i,p)}(Y,S^{**}) \geq \sum_{(i,p)\in\mathcal{R}} \frac{\partial D_{(i,p)}}{\partial s_{vu}}(Y,S^{**})(s^a_{vu}-s^{**}_{vu}) \geq 0,$$

where the last inequality follows from the complementary slackness condition in (24), i.e., $\frac{\partial D_{(i,p)}}{\partial s_{vu}}(Y,S^{**}) = 0$ since $s^{**}_{vu} > 0$, and $\sum_{u\in O_v}\sum_{(i,p):(u,v)\in p} s^{**}_{vu}(i,p) < \hat{s}_v$. Furthermore,

$$\sum_{(i,p)\in\mathcal{R}} D_{(i,p)}(Y,S^a) - \sum_{(i,p)\in\mathcal{R}} D_{(i,p)}(Y,S^{**}) \geq \sum_{(i,p)\in\mathcal{R}} \frac{\partial D_{(i,p)}}{\partial s_{ju}}(Y,S^{**})(s^a_{ju}-s^{**}_{ju}) \geq 0,\ j\in V\backslash v,$$

where the last inequality also follows from the complementary slackness condition in (24). □

If $D_{(i,p)}(Y,S)$ is jointly convex in (Y,S), the above conditions are also sufficient when (25) 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. We do not prove this statement. However, it can be proven using arguments similar to those in [42, Theorem 3].

d) Pareto optimality of $D(Y,S)$: When $f(\mathrm{SINR}_{vu}(S))$ is chosen to be (2), we infer that the sufficiency part of Theorem 2 does not hold since $D_{(i,p)}(Y,S)$ is in general not jointly convex in (Y,S). Hence, we further need to establish the conditions for a Pareto optimal operating point for quasi-convex cost functions (as shown in Prop. 2). We next show that for a solution (Y^{**}, S^{**}) that both satisfies (25) and (26), we have the following Pareto optimal property.

Theorem 3: **Pareto optimality of $D(Y,S)$.** *From Prop. 2 on the quasi-convexity we have $f(\mathrm{SINR}_{vu}(S))$ in (2), $g_{p_k i}(Y)$ in (17), and the relaxed delay-cost function for RCF in (18) are quasi-convex. If a pair of feasible cache and power allocations $((y^{**}_{vi}),(s^{**}_{vu}))$ satisfies conditions (25)-(26) [42, Thm. 3] 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.*

Proof: See Appendix E. □

Given the relaxed delay-cost function $D(Y,S)$ of the form (15), Theorem 3 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 $(i',p')\in\mathcal{R}$.

D. Subgradient Algorithm

For general SINR scenario, the sufficient condition for the convexity given in (20) is not necessarily satisfied, and hence the objective function is not necessarily convex. In this section, we introduce a Clarke subgradient projection method for the non-smooth non-convex problem.

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

Algorithm 2 Projected Subgradient Method

1: Choose S^0, $\boldsymbol{y}^0$, small scalar $\epsilon > 0$ and let $t = 0$
2: **do**
3: Compute Clarke subgradient $d^t_S, d^t_{\boldsymbol{y}}$ by (29);
4: Determine step sizes $\xi^t_{\boldsymbol{y}}$, ξ^t_S according to (31);
5: Compute projected variables $\bar{\boldsymbol{y}}^t$ and $\bar{S}^t$ by (28);
6: Update S^{t+1} and $\boldsymbol{y}^{t+1}$ by (28);
7: Let $t = t+1$;
8: **while** $D^t - D^{t-1} > \epsilon$
9: Let $(\boldsymbol{y}^*_{sub}, S^*_{sub}) = (\boldsymbol{y}^t, S^t)$;
10: Implement *b) Rounding*.

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

$$\begin{aligned} S^{t+1} &= S^t + \xi^t_S(\bar{S}^t - S^t), \quad \bar{S}^t = [S^t - w^t_S d^t_S]^+_{\mathcal{D}_S}, \\ \boldsymbol{y}^{t+1} &= \boldsymbol{y}^t + \xi^t_{\boldsymbol{y}}(\bar{\boldsymbol{y}}^t - \boldsymbol{y}^t), \quad \bar{\boldsymbol{y}}^t = [\boldsymbol{y}^t - w^t_Y d^t_{\boldsymbol{y}}]^+_{\mathcal{D}_{\boldsymbol{y}}}, \end{aligned} \tag{28}$$

where $\xi^t_S, \xi^t_{\boldsymbol{y}} \in (0,1]$ are step sizes respectively corresponding to S and $\boldsymbol{y}$, w^t_S and w^t_Y are positive scalars, $[x]^+_A$ denotes projection of vector x on a convex constraint set A, and

$$d^t_S = \nabla_S D(Y^t,S^t), \quad d^t_{\boldsymbol{y}} \in \partial_{\boldsymbol{y}} D(Y^t,S^t) \tag{29}$$

where d^t_S and $d^t_{\boldsymbol{y}}$ are the Clarke 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 $\boldsymbol{y}$.

2) Subgradient: Provided that the function $D(Y,S)$ is not jointly convex, identifying normal subgradients is generally difficult, as a valid subgradient should form a global underestimator of $D(Y,S)$. Nevertheless, a local generalization of subgradient, i.e., the "Clarke subdifferential",[9] could be calculated efficiently. Specifically, note that since $D(Y,S)$ is continuously differentiable in S over set $\mathcal{D}_S$, the Clarke 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 (16) and using (15), where

$$\partial_{y_{vi}} g_{p_k i} = \begin{cases} \{1\}, & \text{if } \sum_{l=1}^{k} y_{p_l i} < 1 \\ \{0\}, & \text{if } \sum_{l=1}^{k} y_{p_l i} > 1 \\ [0,1], & \text{if } \sum_{l=1}^{k} y_{p_l i} = 1. \end{cases} \tag{30}$$

3) Step Size: The gradient/subgradient magnitudes might be significantly different for Y and S, and therefore we calculate their step sizes separately. We use a modified Polyak's step size [74]. Let $D^t = D(\boldsymbol{y}^t,S^t)$, then

$$\xi^t_{\boldsymbol{y}} = \frac{D^t - \hat{D}^t}{\|d^t_{\boldsymbol{y}}\|^2}, \quad \xi^t_S = \frac{D^t - \hat{D}^t}{\|d^t_S\|^2} \tag{31}$$

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

We summarize the subgradient projection method that achieves the local minima in Algorithm 2.

[9] We refer the readers to [73] for the definition of Clarke subdifferential.

To the best of our knowledge, the theoretical convergence guarantees of the alternating optimization method (Algorithm 1) for non-smooth bi-convex functions, and the Clarke sub-gradient method (Algorithm 2) for non-smooth non-convex functions have not been established.[10] Nevertheless, our extensive simulation results on various network scenarios in Sect. IV demonstrate that both algorithms converge at a desirable rate.

IV. Numerical Results

In this section, we present numerical results for the proposed joint power control and caching network model. We simulate a network in accordance with the model in Sect. II and compare the performance of Algorithms 1 (ALT) and 2 (SUB) to the LRU, LFU and FIFO cache replacement policies for several scenarios, which have been widely used in web caching [76], [77]. These baseline policies depend on a history of requests kept in time, and operate on a time slot basis, whereas our approach is designed to solve a snapshot of the network averaged over time. We pair these baseline 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 7 MCs, where the cell layout is such that one MC at the center and 6 MCs, which model the first-order neighbors of the center MC, surround the centering MC. The network has 5 SCs per MC, and 30 users per MC. Users are distributed uniformly at random within the coverage area of the MC. We use Lloyd's algorithm [78] to construct voronoi cells for this coverage area and place the SCs at the centers of these cells. The users do not cache items, and each one requests a single item at a given time, from a catalog of $|\mathcal{C}| = 150$ items,[11] 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. In our simulations, we capture the interference power from the other MCs, i.e., the inter-cell interference, and incorporate it into the noise level N_u. We also assume that in case of outage events in a given MC, the cost to the backhaul is significantly smaller than the cost of obtaining a neighboring cell. For scenarios with a larger catalog size, our approach can still be implemented conditioning on a selected subset of items based on popularity. For Algorithms 2 and 1, we set the initial points S^0 and Y^0 so that $s^0_{vu} = \hat{s}_v/|O_v|$ and $y^0_{vi} = 0$ for all $v, u \in V$ and $i \in \mathcal{C}$. While the algorithms we propose 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 our algorithms.

We next detail our observations from five distinct simulation settings. To measure the gains, the default setting for simulations is that the SC cache capacity is $c_{sc} = 5$, the MC cache capacity is $c_{mc} = 10$, $\hat{s}_v = 200$, $\gamma = 0.5$, the path loss exponent is $n = 3.7$ and the noise power is $N_u = 1$ for all $u \in V$. Below, in each figure, we plot D^o over one of these parameters while keeping the others as default. We include other necessary parameters and details in these discussions.

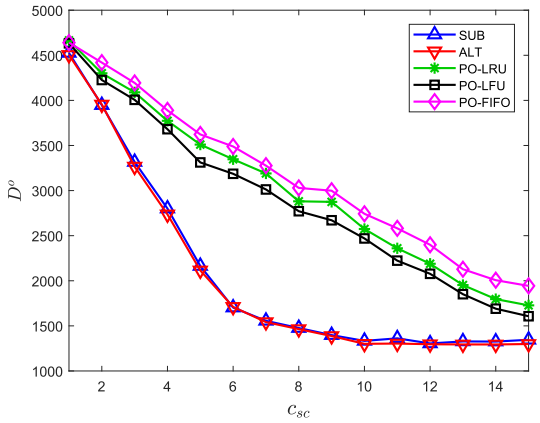


Fig. 2. D^o versus increasing small cell cache capacity (assume $c_{mc} = 2\ c_{sc}$), with $\gamma = 0.5$, $\hat{s}_v = 200$ and $N_u = 1$.

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 algorithms reduce delay at a much faster rate compared to traditional replacement algorithms. SUB and ALT algorithms also achieve a point of minimum delay given large enough caches, while traditional algorithms do not converge to such a point and perform worse than SUB and ALT with all values of the cache capacity constraint. Numerically, SUB and ALT methods achieve up to 50% less delay at $c_{sc} = 6$ compared with traditional algorithms. The improvement begins to saturate at $c_{sc} = 6$ and diminishes at $c_{sc} = 10$. This is because, with a large enough cache capacity, almost all requests are fulfilled at a nearby SC, and we can hardly improve the delay by providing more caches. The simulation result also implies that for delay saturation, it takes up to $2\times$ the cache capacity for traditional algorithms to match the delay performance of SUB and ALT.

Effect of power constraints. We present the results of this setting in Fig. 3. We observe that traditional methods and our algorithms 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 algorithms achieve up to 50% less delay versus the best-performing traditional method, POLFU. Moreover, when the individual power budget $\hat{s}_v$ is more than 100, the decrease speed of total delay versus power budget drops to a low level. This result is because when the power budget is high enough, the noise (combined with the inter-cell interference) is negligible compared with intra-cell interference, which limits the SINR. A higher power budget would not help reduce the average intra-cell interference level.

Effect of request distribution. We sketch the delay behavior as a function of the Zipf exponent γ in Fig. 4. As γ increases, the requests become more skewed towards the most popular items, which causes a reduction in the delay. We also observe the delay performance of SUB and ALT methods is up to 30% better than traditional replacement models reinforced with

[10] We refer the readers to [71] for bi-convex optimization and [75] for Clarke-subgradient optimization.

[11] We note that the choice of catalog size is aligned with the literature on geographic caching, e.g., in [15], where $|\mathcal{C}| = 25$ and cache memories are of size 5, and more recently in [79] and [80] for $|\mathcal{C}|$ ranging from 100 to 300.

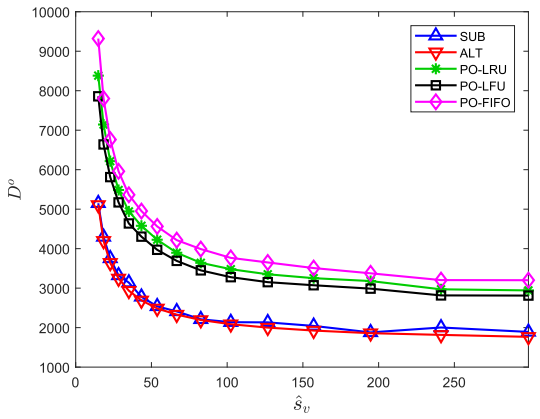


Fig. 3. D^o versus node power budget $\hat{s}_v$, with $c_{sc} = 5$, $c_{mc} = 10$, $\gamma = 0.5$ and $N_u = 1$.

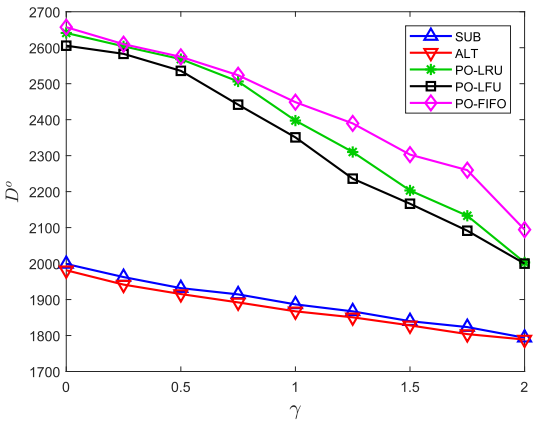


Fig. 4. D^o versus parameter γ of Zipf distribution of requested items, with $c_{sc} = 5$, $c_{mc} = 10$, $\hat{s}_v = 200$ and $N_u = 1$.

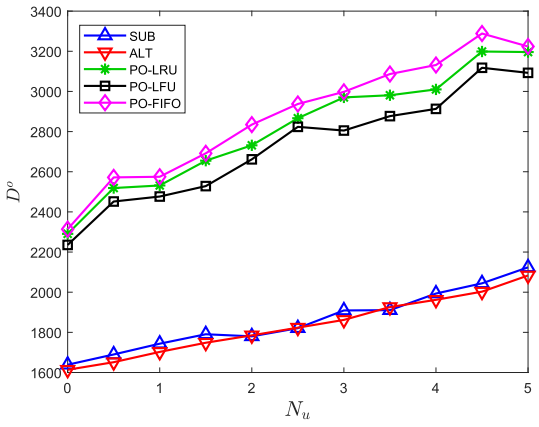


Fig. 5. D^o versus receiver noise power N_u, with $c_{sc} = 5$, $c_{mc} = 10$, $\hat{s}_v = 200$ and $\gamma = 0.5$.

power optimization. The advantage of SUB and ALT against the other algorithms decreases with γ because as γ increases, the Zipf distribution becomes more centralized to the popular content items, reducing the size of popular items. Thus, with increasing γ, the advantage of SUB and ALT against the traditional priority-based caching policies diminishes.

Effect of noise power. We capture the effect of the noise N_u at the receiving user of the centering MC. As shown in Fig. 5, the delay of all algorithms degrades with the increasing noise because the effective SINR of each link degrades. The delay performance of SUB and ALT is up to 50% better than traditional algorithms across different noise levels N_u.

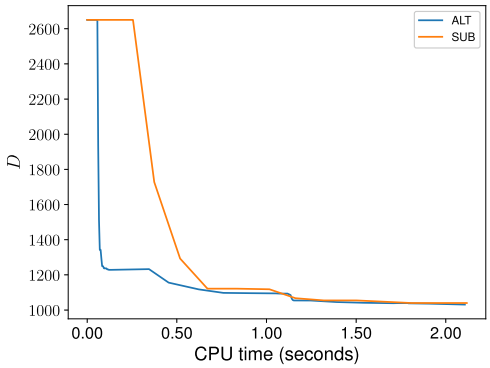


Fig. 6. Convergence of SUB and ALT algorithms as described by D with respect to time. $c_{sc} = 2$, $c_{mc} = 4$, $\hat{s}_v = 100$ and $\gamma = 0.25$.

Convergence of SUB and ALT. We present the results of this setting in Fig. 6. We observe that while both algorithms reach the minimum in similar times, ALT has a much steeper initial decrease in the relaxed delay-cost D. This is because the number of power variables is significantly smaller than the number of caching variables. ALT optimizes them separately which results in the initial steep decrease where power variables are being optimized whereas SUB optimizes them jointly leading to longer durations for each iteration. Furthermore, the subhessian matrix of $D(Y, S)$ does not satisfy the necessary properties for joint convexity, as we detailed in Sect. III-A.3, causing a slow convergence for SUB.

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. More specifically, we provided necessary and sufficient conditions for the optimality of RCF to the joint optimization problem. We demonstrated Pareto optimality of the solution to RCF and devised two solution techniques: alternating optimization technique when RCF satisfies biconvexity and subgradient projection algorithm for general non-convex RCF. The results of our approach can enable the wireless HetNets to optimally exploit the resources to minimize the use of the backhaul connection, hence reduce the transmission delays in both mobile devices and the infrastructure, and 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.

Possible extensions of this work include devising decentralized techniques and designing both uncoded and coded caching schemes via distributed adaptive stochastic descent algorithms and implementing them in practice. Furthermore, these results will inform the design of edge cloud architectures by clarifying the relative benefits of centralized and distributed implementation. Querying for content can be seen as a simplified case of querying for a result of a computation or service. Thus, caching and routing algorithms are essential ingredients of an edge computing infrastructure that optimally schedules processing and job flows. Therefore, quantifying the potential cost savings from the deployment of SCs with

caching capabilities via optimal routing algorithms is critical. Extensions also include leveraging more practical interference mitigation schemes instead of having all transmissions on the same frequency, and a more detailed analysis of backhaul costs to effectively route the requests and control the traffic load on SCs and to overcome the transmission delay incurred in the backhaul due to limited bandwidth and dynamic channel conditions. Further incorporating the unequal file sizes and the (non-instantaneous) cache download times will improve the accuracy of the proposed framework. Finally, our proposed framework could also be extended to incorporate energy efficiency in wireless HetNets, by adding a term of weighted total power consumption in the objective.

Appendix A
Proof of Theorem 1

The proof follows from relaxing and bounding techniques. By Goemans and Williamson [62], [66], we have

$$\prod_{l=1}^{k}(1-y_{p_l i}) \le 1-(1-(1-1/k)^k)\min\Big\{1,\sum_{l=1}^{k} y_{p_l i}\Big\} \le 1-(1-1/e)\min\Big\{1,\sum_{l=1}^{k} y_{p_l i}\Big\}, \tag{32}$$

as $(1-1/k)^k \le 1/e$. On the other hand, we have $D^o(Y,S) = \mathbb{E}_\nu[D^o(X,S)]$. Using (32), the relaxed cost function for serving a request (given that each request $(i,p)\in\mathcal{R}$ is well-routed) can be written concisely in terms of the allocation:

$$D(Y,S) \overset{(a)}{\ge} \sum_{(i,p)\in\mathcal{R}} \lambda_{(i,p)} \sum_{k=1}^{|p|-1} f(\mathrm{SINR}_{p_{k+1}p_k}(S)) \prod_{l=1}^{k}(1-y_{p_l i}) = D^o(Y,S), \tag{33}$$

where (a) is due to (32). We also upper bound $D(Y,S)$ as

$$D(Y,S) \overset{(b)}{\le} \sum_{(i,p)\in\mathcal{R}} \lambda_{(i,p)} \sum_{k=1}^{|p|-1} f(\mathrm{SINR}_{p_{k+1}p_k}(S))\mathbb{E}_\nu\left[g_{p_k i}(X)\right], \tag{34}$$

where (b) follows from the concavity of the min operator and employing (17), and

$$\begin{aligned}\mathbb{E}_\nu\left[g_{p_k i}(X)\right] &\overset{(c)}{\le} 1-(1-1/e)\mathbb{E}_\nu\Big[\min\Big\{1,\sum_{l=1}^{k} x_{p_l i}\Big\}\Big] \\ &= 1-(1-1/e)\mathbb{E}_\nu\Big[1-\prod_{l=1}^{k} x_{p_l i}\Big] \\ &= 1-(1-1/e)\Big(1-\prod_{l=1}^{k} y_{p_l i}\Big),\end{aligned}$$

where (c) is due to $(1-1/k)^k \le 1/e$. Hence, $D^o(Y,S) \ge (D(Y,S)-D^{\mathrm{ub}}/e)/(1-1/e)$. Hence,

$$D(Y,S) \le D^{\mathrm{ub}}/e + (1-1/e)D^o(Y,S), \tag{35}$$

From (33) and (34), we have

$$D(Y,S) \ge D^o(Y,S) \ge (D(Y,S)-D^{\mathrm{ub}}/e)/(1-1/e). \tag{36}$$

Because $Y^* \in \mathcal{D}_Y$ is optimal, $D^o(Y^*,S) \le D^o(Y^{**},S)$. From (36) and the optimality of Y^{**}, $D^o(Y^{**},S) \le D(Y^{**},S) \le D(Y^*,S) \le D^{\mathrm{ub}}/e + (1-1/e)D^o(Y^*,S)$.

Appendix B
Proof of Proposition 1

Note that the objective function $D(S,Y)$ in (15) is convex in caching variables Y. It is convex in S if every $f(\mathrm{SINR}_{p_{k+1}p_k}(S))$ is convex in S where $\mathrm{SINR}_{p_{k+1}p_k}(S)$ is concave in S for all k. However, given that $\mathrm{SINR}_{p_{k+1}p_k}$ is strictly increasing, $\nabla^2\mathrm{SINR}_{p_{k+1}p_k}(S)$ cannot be negative definite. Letting $C = f^{-1}$, based on the observations [69], if

$$C''(x)\cdot x + C'(x) \le 0, \quad \forall x \ge 0, \tag{37}$$

then C is concave in $P \triangleq (\log(s_{vu}(i,p)))_{(v,u)\in E,\,(i,p):(u,v)\in p}$ (power measured in dB). From above relation, since $C''(x) = \frac{2f'(x)^2 - f(x)f''(x)}{f(x)^3} \le 0$, and $f(x) \ge 0$, $\forall x \ge 0$, we have $2f'(x)^2 - f(x)f''(x) \le 0$, yielding $0 \le \frac{2f'(x)^2}{f(x)} \le f''(x)$. Hence, $f(\mathrm{SINR}_{p_{k+1}p_k})$ is convex in P.

The condition in (37) equivalently yields the following for $f^{-1}(\mathrm{SINR}_{p_{k+1}p_k})$ to be concave in P:

$$\frac{2f'(x)^2 - f(x)f''(x)}{f(x)^3}\cdot x - \frac{f'(x)}{f(x)^2} \le 0. \tag{38}$$

Reordering the terms in (38) we get the desired result.

In addition to the condition in (38) on convexity in log power variables, $f(\mathrm{SINR}_{vu}(S)$ might be convex in s_{vu} when the nodes have a total power constraint as given by (9) which is satisfied with equality. We next explain the requirement under which the convexity in powers holds. Assuming that the total transmit power is fixed and equal to $\hat{s}$, since the routings are predetermined, a user's total received power lumped with the noise power, coined $\bar{s}$, will be fixed given the allocation of the transmit power. Hence, $f(\mathrm{SINR}_{vu}(S) = \frac{1}{\log_2(1+\frac{s_{vu}}{\bar{s}-s_{vu}})}$ for $(v,u)\in E$. Using this relation and computing the second order derivative of $f(\mathrm{SINR}_{vu}(S)$ with respect to s_{vu}, it can be shown via algebraic manipulation that $f(\mathrm{SINR}_{vu}(S)$ is convex in s_{vu} provided that $s_{vu} \le \frac{3}{4}\bar{s}$. We also emphasize that this condition is not restricted to the high SINR regime and is valid under general SINR values, provided that the total transmission power constraint is satisfied with equality.

Appendix C
Proof of Proposition 2

With no loss of generality, assume that $S^a \ne S^b$ such that $D(Y,S^a) < D(Y,S^b)$. Then, a function D defined on a convex subset $\mathcal{D}_Y \times \mathcal{D}_S$ of a real vector space is quasi-convex in S given Y, if for all $S^a \ne S^b$ and $\alpha \in (0,1)$ we have the following condition:

$$D(Y,\alpha S^a + (1-\alpha)S^b) < D(Y,S^b). \tag{39}$$

Since D is linear in f, it is easily verified that a sufficient condition for $D(Y,S)$ to be quasi-convex in S is when f is quasi-convex. While the sum of quasiconvex functions defined on the same domain need not be quasiconvex, we will detail why quasi-convexity will be preserved in our setting.

For $f(\mathrm{SINR}_{vu}(S^a)) < f(\mathrm{SINR}_{vu}(S^b))$ we have

$$\begin{aligned} &f(\mathrm{SINR}_{vu}(\alpha S^a + (1-\alpha)S^b)) \\ &= \frac{1}{\log_2(1+\mathrm{SINR}_{vu}(\alpha S^a + (1-\alpha)S^b))} \\ &< \frac{1}{\log_2(1+\mathrm{SINR}_{vu}(S^b))} = f(\mathrm{SINR}_{vu}(S^b)),\end{aligned} \tag{40}$$

where we used (2). Under a total power constraint we note that $f(\mathrm{SINR}_{vu}(S))$ decreases in s_{vu}. Since $f(\mathrm{SINR}_{vu}(\alpha S^a + (1-\alpha)S^b)) < \max\left\{f(\mathrm{SINR}_{vu}(S^a)), f(\mathrm{SINR}_{vu}(S^b))\right\} = f(\mathrm{SINR}_{vu}(S^b))$ for all $(v,u) \in E$ and max is Schur-convex, and by ordering the summands of $\sum_{k=1}^{|p|-1} f(\mathrm{SINR}_{p_{k+1}p_k}(S))$ in a decreasing manner, we have that

$$\max\Big\{\sum_{k=1}^{|p|-1} f(\mathrm{SINR}_{p_{k+1}p_k}(S^a)), \sum_{k=1}^{|p|-1} f(\mathrm{SINR}_{p_{k+1}p_k}(S^b))\Big\} = \sum_{k=1}^{|p|-1} f(\mathrm{SINR}_{p_{k+1}p_k}(S^b)).$$

Hence, we infer from (40) and the order-preserving mapping that (39) holds which implies $D(Y,S^a) < D(Y,S^b)$.

We can also observe that

$$\begin{aligned} &f(\alpha\mathrm{SINR}_{vu}(S^a) + (1-\alpha)\mathrm{SINR}_{vu}(S^b)) \\ &\quad = \frac{1}{\log_2(1+\alpha\mathrm{SINR}_{vu}(S^a) + (1-\alpha)\mathrm{SINR}_{vu}(S^b))} \\ &\quad < \frac{1}{\log_2(1+\mathrm{SINR}_{vu}(S^b))} = f(\mathrm{SINR}_{vu}(S^b)), \end{aligned} \tag{41}$$

where inequality follows from that $\mathrm{SINR}_{vu}(S^a) > \mathrm{SINR}_{vu}(S^b)$ which implies $\alpha\mathrm{SINR}_{vu}(S^a) + (1-\alpha)\mathrm{SINR}_{vu}(S^b) > \mathrm{SINR}_{vu}(S^b)$. Furthermore, $f(\mathrm{SINR}_{vu}(S))$ is a monotonically decreasing function of $\mathrm{SINR}_{vu}(S)$. Hence, $f(\mathrm{SINR}_{vu}(S))$ is quasi-convex in $\mathrm{SINR}_{vu}(S)$.

Note that $D(Y,S)$ is convex with respect to set $\mathcal{D}_Y$, which is due to (17). Note also that D is quasi-convex in Y. This can be shown using the condition that if

$$D(Y^a,S) < D(Y^b,S), \tag{42}$$

then for any $\lambda \in (0,1)$ it holds that

$$D(\lambda Y^a + (1-\lambda)Y^b, S) < D(Y^b,S). \tag{43}$$

To verify D is quasi-convex, it is necessary that $g_{p_k i}(Y)$'s given in (17) are quasi-convex. Assume $g_{p_k i}(Y^a) < g_{p_k i}(Y^b)$. Then for all $Y^a \neq Y^b$ and $\lambda \in (0,1)$:

$$\begin{aligned} &g_{p_k i}(\lambda Y^a + (1-\lambda)Y^b) \\ &= 1 - a_k \min\Big\{1, \sum_{l=1}^{k} \lambda y^a_{p_l i} + (1-\lambda) y^b_{p_l i}\Big\} \\ &\overset{(a)}{\leq} 1 - a_k\Big[\lambda \min\Big\{1, \sum_{l=1}^{k} y^a_{p_l i}\Big\} + (1-\lambda)\min\Big\{1, \sum_{l=1}^{k} y^b_{p_l i}\Big\}\Big] \\ &= \lambda g_{p_k i}(Y^a) + (1-\lambda) g_{p_k i}(Y^b) < g_{p_k i}(Y^b), \end{aligned} \tag{44}$$

where (a) is due to the concavity of the min function. This verifies that D is quasi-convex in Y.

Appendix D
Proof of Theorem 2

If the optimization problem satisfies some regularity conditions [65], the necessary conditions for a solution of the nonlinear RCF in (18) are given by the KKT conditions. Then, the following four groups of conditions hold:

(i) For stationary, the local solution (Y^{**}, S^{**}) needs to satisfy the following subgradients with respect to $Y = (y_{vi})$ and the gradients with respect to $S = (s_{vu})$ values:

$$d_{y_{vi}} D(Y^{**},S) + \sum_{v,i}\mu_{vi} - \sum_{v,i}\nu_{vi} + \sum_{v\in V}\eta_v = 0, \quad \forall\, v,i,$$

$$\frac{\partial D(Y,S^{**})}{\partial s_{vu}} + \sum_{v\in V}\beta_v - \sum_{e\in E,\, r\in\mathcal{R}}\gamma_{er} = 0, \quad \forall (u,v),(i,p).$$

(ii) For primal feasibility, we require that

$$y^{**}_{vi} - 1 \leq 0, \quad v\in V,\ i\in\mathcal{C}, \tag{45}$$

$$-y^{**}_{vi} \leq 0, \quad v\in V,\ i\in\mathcal{C}, \tag{46}$$

$$\sum_{i\in\mathcal{C}} y_{vi} \leq c_v, \quad v\in V, \tag{47}$$

$$\sum_{u\in O_v} s^{**}_{vu} - \hat{s}_v \leq 0, \quad v\in V, \tag{48}$$

$$-s^{**}_{vu} \leq 0, \quad v,\, u\in V. \tag{49}$$

For dual feasibility, we require that

$$\begin{aligned} &\mu_{vi},\ \nu_{vi},\ \eta_v,\ \beta_v \geq 0, \quad v\in V,\ i\in\mathcal{C}, \\ &\gamma_{e,r} \geq 0, \quad e=(u,v)\in E,\ r=(i,p)\in\mathcal{R} \end{aligned} \tag{50}$$

(iii) The complementary slackness conditions are given as

$$\begin{aligned} \mu_{vi}(y^{**}_{vi} - 1) &= 0, \quad v\in V,\ i\in\mathcal{C}, \\ \nu_{vi}\cdot y^{**}_{vi} &= 0, \quad v\in V,\ i\in\mathcal{C}, \\ \eta_v\Big(\sum_{i\in\mathcal{C}} y_{vi} - c_v\Big) &= 0, \quad v\in V, \\ \beta_v\Big(\sum_{u\in O_v} s^{**}_{vu} - \hat{s}_v\Big) &= 0, \quad v\in V, \\ \gamma_{e,r}\cdot s^{**}_{vu} &= 0, \quad e=(u,v)\in E,\ r\in\mathcal{R}. \end{aligned} \tag{51}$$

(iv) The subgradients with respect to y_{vi} should satisfy

$$\begin{aligned} &d_{y_{vi}} D(Y^{**},S) \\ &= \sum_{u\in O_v}\sum_{(i,p):(u,v)\in p}\lambda_{(i,p)}\sum_{k=1}^{|p|-1} f(\mathrm{SINR}_{p_{k+1}p_k}(S)) d_{y_{vi}} g_{p_k i}(Y^{**}) \\ &= -\mu_{vi} + \nu_{vi} + \eta_v = \alpha_{vi}, \quad v\in V,\ i\in\mathcal{C}. \end{aligned} \tag{52}$$

When $y_{vi} = 0$, constraint (45) is eliminated, and $d_{y_{vi}} D(Y^{**},S) \geq \alpha_{vi}$. Similarly, if $y_{vi} = 1$, constraint (46) is eliminated, and $d_{y_{vi}} D(Y^{**},S) < \alpha_{vi}$. This verifies (23).

(v) The gradients with respect to the power variables are

$$\begin{aligned} &\frac{\partial D(Y,S^{**})}{\partial s_{vu}} = \sum_{u\in O_v}\sum_{(i,p):(u,v)\in p}\lambda_{(i,p)}\frac{\partial f(\mathrm{SINR}_{vu}(S^{**}))}{\partial s_{vu}} g_{ui}(Y) \\ &= -\beta_v + \gamma_{e,r}, \quad e=(u,v)\in E,\ r=(i,p)\in\mathcal{R}. \end{aligned} \tag{53}$$

Solving the gradients (52) and (53), along with the complementary slackness conditions in (51), we obtain the necessary conditions, which concludes the proof.

Appendix E
Proof of Theorem 3

The proof follows from employing similar techniques as in [42, Sect. VI-C, Theorem 4].

We initially assume that the joint power and cache allocation problem is quasi-convex. With this assumption, for a fixed cache allocation $(y^{**}_{vi})_{v\in V,\, i\in\mathcal{C}}$, the relaxed delay-cost is a

convex function of S. Therefore, any feasible power allocation S^* satisfying (26) satisfies that

$$D(Y, S^{**}) = \min_{S \in \mathcal{D}_S} \sum_{(i,p)\in\mathcal{R}} D_{(i,p)}(Y, S). \tag{54}$$

Given any feasible power allocation, if condition (25) holds at cache allocation $(y_{vi}^{**})_{v\in V,\, i\in\mathcal{C}}$, then

$$D(Y^{**}, S) = \min_{Y \in \mathcal{D}_Y} \sum_{(i,p)\in\mathcal{R}} D_{(i,p)}(Y, S). \tag{55}$$

In this case, any initial power and cache allocation configuration can be driven to a limiting (Y^{**}, S^{**}) such that the condition (26) is satisfied at S^{**} given Y^{**}, and Y^{**} satisfies (25) given S^{**}. We now suppose that under the more general convex power allocation region model, there are algorithms that also can drive the power and cache configuration to a limit (Y^{**}, S^{**}) such that the conditions (26) and (25) hold simultaneously. Although global optimality cannot be guaranteed, the Pareto optimality can be shown.

Suppose that $D(Y^{\#}, S^{\#})$ Pareto dominates $D(Y^{**}, S^{**})$. Without loss of generality, we can assume

$$D_{(m,r)}(Y^{\#}, S^{\#}) < D_{(m,r)}(Y^{**}, S^{**}). \tag{56}$$

Because both $Y^{\#}$ and Y^{**} belong to $\mathcal{D}_Y$, and $\mathcal{D}_Y$ is strictly convex, $Y^{\beta} = \beta Y^{**} + (1-\beta)Y^{\#}$ is achievable for all $\beta \in (0,1)$. Moreover, $Y^{\#} \neq Y^{**}$ because otherwise, $S^{\#} \neq S^{**}$, and from Pareto domination we would have

$$\sum_{(i,p)\in\mathcal{R}} D_{(i,p)}(Y^{**}, S^{\#}) < \sum_{(i,p)\in\mathcal{R}} D_{(i,p)}(Y^{**}, S^{**}). \tag{57}$$

However, this contradicts (54). Therefore, Y^{β} is in the interior of $\mathcal{D}_Y$ for any $\beta \in (0,1)$.

From the same reasoning, $S^{\#} \neq S^{**}$ and $S^{\alpha} = \alpha S^{**} + (1-\alpha)S^{\#}$ is feasible for any $\alpha \in [0,1]$ simply by linearity of feasible power allocations.

Since $D_{(i,p)}$ is quasi-convex, $D(Y^{\alpha}, S^{\alpha})$ Pareto dominates $D(Y^{**}, S^{**})$ as well for any $\alpha \in (0,1)$, since $D_{(m,r)}(Y^{\alpha}, S^{\alpha}) \leq \max\{D_{(m,r)}(Y^{\#}, S^{\alpha}),\, D_{(m,r)}(Y^{**}, S^{\alpha})\}$ $< D_{(m,r)}(Y^{**}, S^{**})$, and $D_{(i,p)}(Y^{\alpha}, S^{\alpha}) \leq D_{(i,p)}(Y^{**}, S^{**})$ for $(i,p) \neq (m,r)$. Summing up all the terms on LHS and RHS, we have for any $\alpha \in (0,1)$

$$\sum_{(i,p)\in\mathcal{R}} D_{(i,p)}(Y^{\alpha}, S^{\alpha}) < \sum_{(i,p)\in\mathcal{R}} D_{(i,p)}(Y^{**}, S^{**}). \tag{58}$$

By optimality condition (26) and the fact that Y^{α} is in the interior of $\mathcal{D}_Y$ for any $\alpha \in (0,1)$, we have for any $\alpha \in (0,1)$ and $(v,u) \in E$,

$$\begin{aligned}
&\sum_{(i,p)\in\mathcal{R}} \frac{\partial D_{(i,p)}}{\partial s_{vu}}(Y^{**}, S^{**})(S^{\#} - S^{**}) \\
&\quad \overset{(a)}{=} \frac{1}{1-\alpha} \sum_{(i,p)\in\mathcal{R}} \frac{\partial D_{(i,p)}}{\partial s_{vu}}(Y^{**}, S^{**})(S^{\alpha} - S^{**}) \\
&\quad > \frac{1}{1-\alpha} \sum_{(i,p)\in\mathcal{R}} \frac{\partial D_{(i,p)}}{\partial s_{vu}}(Y^{**}, S^{**})(\bar{S}^{\alpha} - S^{**}) \geq 0,
\end{aligned}$$

where (a) follows from $S^{\alpha} = \alpha S^{**} + (1-\alpha)S^{\#}$, and $\bar{S}^{\alpha}$ is some power matrix strictly dominating S^{α}. Following similar steps, from (27) for $j \in V\backslash v,\ (i',p') : (u,j) \in p'$

$$\begin{aligned}
&\sum_{(i,p)\in\mathcal{R}} \frac{\partial D_{(i,p)}}{\partial s_{ju}}(Y^{**}, S^{**})(S^{\#} - S^{**}) \\
&\quad > \frac{1}{1-\alpha} \sum_{(i,p)\in\mathcal{R}} \frac{\partial D_{(i,p)}}{\partial s_{ju}}(Y^{**}, S^{**})(\bar{S}^{\alpha} - S^{**}) \geq 0.
\end{aligned}$$

Since $D_{(i,p)}$ is twice continuously differentiable on S, there exists $\epsilon > 0$ such that for all $\alpha \in [1-\epsilon, 1)$

$$\sum_{(i,p)\in\mathcal{R}} \frac{\partial D_{(i,p)}}{\partial s_{vu}}(Y^{\alpha}, S^{**})(S^{\alpha} - S^{**}) \geq 0, \tag{59}$$

$$\sum_{(i,p)\in\mathcal{R}} \frac{\partial D_{(i,p)}}{\partial s_{ju}}(Y^{\alpha}, S^{**})(S^{\alpha} - S^{**}) \geq 0,$$
$$j \in V\backslash v,\ (i',p') : (u,j) \in p'. \tag{60}$$

Combining (59) with the convexity of $D_{(i,p)}(Y^{\alpha}, \cdot)$ implies

$$\begin{aligned}
\sum_{(i,p)\in\mathcal{R}} D_{(i,p)}(Y^{\alpha}, S^{**}) &\leq \sum_{(i,p)\in\mathcal{R}} D_{(i,p)}(Y^{\alpha}, S^{\alpha}) \\
&< \sum_{(i,p)\in\mathcal{R}} D_{(i,p)}(Y^{**}, S^{**})
\end{aligned} \tag{61}$$

where the second inequality comes from (58). However, this result contradicts (55). Hence, 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.

REFERENCES

[1] J. G. Andrews et al., "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. 8402–8413, Dec. 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. 1305–1314, Sep. 2002.

[4] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and Zipf-like distributions: Evidence and implications," in *Proc. IEEE INFOCOM*, Mar. 1999, vol. 1, no. 1, New York, NY, USA, pp. 126–134.

[5] M. Garetto, E. Leonardi, and V. Martina, "A unified approach to the performance analysis of caching systems," *ACM Trans. Model. Perform. Eval. Comput. Syst.*, vol. 1, no. 3, p. 12, 2016.

[6] E. Cohen and S. Shenker, "Replication strategies in unstructured peer-to-peer networks," in *Proc. Conf. Appl., Technol., Architectures, Protocols Comput. Commun.*, Pittsburgh, PA, USA, Aug. 2002, pp. 177–190.

[7] A. Dan and D. Towsley, "An approximate analysis of the LRU and FIFO buffer replacement schemes," in *Proc. ACM SIGMETRICS Conf. Meas. Model. Comput. Syst.*, 1990, pp. 143–152.

[8] D. S. Berger, P. Gland, S. Singla, and F. Ciucu, "Exact analysis of TTL cache networks," *Perform. Eval.*, vol. 79, pp. 2–23, Sep. 2014.

[9] M. Leconte, G. Paschos, L. Gkatzikis, M. Draief, S. Vassilaras, and S. Chouvardas, "Placing dynamic content in caches with small population," in *Proc. IEEE INFOCOM 35th Annu. IEEE Int. Conf. Comput. Commun.*, San Francisco, CA, USA, Apr. 2016, pp. 1–9.

[10] S. Traverso, M. Ahmed, M. Garetto, P. Giaccone, E. Leonardi, and S. Niccolini, "Temporal locality in today's content caching: Why it matters and how to model it," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 43, no. 5, pp. 5–12, Nov. 2013.

[11] E. Leonardi and G. L. Torrisi, "Least recently used caches under the shot noise model," in *Proc. IEEE Conf. Comput. Commun. (INFOCOM)*, Hong Kong, Apr. 2015, pp. 2281–2289.

[12] C. Fricker, P. Robert, and J. Roberts, "A versatile and accurate approximation for LRU cache performance," in *Proc. 24th Int. Teletraffic Congr.*, Anaheim, CA, USA, Sep. 2012, pp. 1–18.

[13] S. Ioannidis and E. Yeh, "Adaptive caching networks with optimality guarantees," in *Proc. ACM SIGMETRICS Int. Conf. Meas. Model. Comput. Sci.*, Jun. 2016, pp. 113–124.

[14] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.

[15] B. Blaszczyszyn and A. Giovanidis, "Optimal geographic caching in cellular networks," in *Proc. IEEE Int. Conf. Commun.*, London, U.K., Jun. 2015, pp. 3358–3363.

[16] N. Golrezaei, P. Mansourifard, A. F. Molisch, and A. G. Dimakis, "Base-station assisted device-to-device communications for high-throughput wireless video networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 7, pp. 3665–3676, Jul. 2014.

[17] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," *IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 849–869, Dec. 2015.

[18] D. Malak, M. Al-Shalash, and J. G. Andrews, "Spatially correlated content caching for device-to-device communications," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 56–70, Jan. 2018.

[19] U. Niesen, D. Shah, and G. W. Wornell, "Caching in wireless networks," *IEEE Trans. Inf. Theory*, vol. 58, no. 10, pp. 6524–6540, Oct. 2012.

[20] M. Ji, G. Caire, and A. F. Molisch, "The throughput-outage tradeoff of wireless one-hop caching networks," *IEEE Trans. Inf. Theory*, vol. 61, no. 12, pp. 6833–6859, Dec. 2015.

[21] L. Zhang, M. Xiao, G. Wu, and S. Li, "Efficient scheduling and power allocation for D2D-assisted wireless caching networks," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2438–2452, Jun. 2016.

[22] D. Malak, M. Al-Shalash, and J. G. Andrews, "Optimizing content caching to maximize the density of successful receptions in device-to-device networking," *IEEE Trans. Commun.*, vol. 64, no. 10, pp. 4365–4380, Oct. 2016.

[23] S.-W. Jeon, S.-N. Hong, M. Ji, and G. Caire, "Caching in wireless multihop device-to-device networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, London, U.K., Jun. 2015, pp. 6732–6737.

[24] S. Ioannidis and E. Yeh, "Jointly optimal routing and caching for arbitrary network topologies," in *Proc. 4th ACM Conf. Inf.-Centric Netw.*, Sep. 2017, pp. 77–87.

[25] M. Dehghan, L. Massoulié, D. Towsley, D. S. Menasché, and Y. C. Tay, "A utility optimization approach to network cache design," *IEEE/ACM Trans. Netw.*, vol. 27, no. 3, pp. 1013–1027, Jun. 2019.

[26] E. Yeh, T. Ho, Y. Cui, M. Burd, R. Liu, and D. Leong, "VIP: A framework for joint dynamic forwarding and caching in named data networks," in *Proc. 1st ACM Conf. Inf.-Centric Netw.*, Paris, France, Sep. 2014, pp. 117–126.

[27] 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, USA, May 2018, pp. 1–7.

[28] I. Baev, R. Rajaraman, and C. Swamy, "Approximation algorithms for data placement problems," *SIAM J. Comput.*, vol. 38, no. 4, pp. 1411–1429, Aug. 2008.

[29] J. Li et al., "DR-cache: Distributed resilient caching with latency guarantees," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, Honolulu, HI, USA, Apr. 2018, pp. 441–449.

[30] M. Dehghan et al., "On the complexity of optimal routing and content caching in heterogeneous networks," in *Proc. IEEE INFOCOM*, Hong Kong, Apr. 2015, pp. 936–944.

[31] 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, Jun. 2014.

[32] Z. Chen, N. Pappas, and M. Kountouris, "Probabilistic caching in wireless D2D networks: Cache hit optimal versus throughput optimal," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 584–587, Mar. 2017.

[33] B. Blaszczyszyn, P. Keeler, and P. Muhlethaler, "Optimizing spatial throughput in device-to-device networks," in *Proc. 15th Int. Symp. Modeling Optim. Mobile, Ad Hoc, Wireless Netw.*, Paris, France, May 2017, pp. 1–6.

[34] M. Xiao, N. B. Shroff, and E. K. P. Chong, "A utility-based power-control scheme in wireless cellular systems," *IEEE/ACM Trans. Netw.*, vol. 11, no. 2, pp. 210–221, Apr. 2003.

[35] H. Gupta, N. He, and R. Srikant, "Optimization and learning algorithms for stochastic and adversarial power control," in *Proc. Int. Symp. Model. Optim. Mobile, Ad Hoc, Wireless Netw. (WiOPT)*, Paris, France, Jun. 2019, pp. 1–8.

[36] S. V. Hanly, "An algorithm for combined cell-site selection and power control to maximize cellular spread spectrum capacity," *IEEE J. Sel. Areas Commun.*, vol. 13, no. 7, pp. 1332–1340, Sep. 1995.

[37] S. V. Hanly and D.-N. Tse, "Power control and capacity of spread spectrum wireless networks," *Automatica*, vol. 35, no. 12, pp. 1987–2012, Dec. 1999.

[38] S. V. Hanly, "Capacity and power control in spread spectrum macrodiversity radio networks," *IEEE Trans. Commun.*, vol. 44, no. 2, pp. 247–256, Feb. 1996.

[39] C. Liaskos, X. Dimitropoulos, and L. Tassiulas, "Backpressure on the backbone: A lightweight, non-intrusive traffic engineering approach," *IEEE Trans. Netw. Service Manag.*, vol. 14, no. 1, pp. 176–190, Mar. 2017.

[40] A. Dvir and N. Carlsson, "Power-aware recovery for geographic routing," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Budapest, Hungary, Apr. 2009, pp. 2851–2856.

[41] J.-H. Chang and L. Tassiulas, "Energy conserving routing in wireless ad-hoc networks," in *Proc. IEEE Int. Conf. Comput. Commun.*, Mar. 2000, pp. 22–31.

[42] 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.

[43] O. Chipara et al., "Real-time power-aware routing in sensor networks," in *Proc. 14th IEEE Int. Workshop Quality Service*, Jun. 2006, pp. 83–92.

[44] S. Barbarossa, S. Sardellitti, and P. Di Lorenzo, "Joint allocation of computation and communication resources in multiuser mobile cloud computing," in *Proc. IEEE 14th Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Jun. 2013, pp. 26–30.

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

[46] J. Liu, Y. Mao, J. Zhang, and K. B. Letaief, "Delay-optimal computation task scheduling for mobile-edge computing systems," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Barcelona, Spain, Jul. 2016, pp. 1451–1455.

[47] R. Gallager, "A minimum delay routing algorithm using distributed computation," *IEEE Trans. Commun.*, vol. COM-25, no. 1, pp. 73–85, Jan. 1977.

[48] J. Oueis, E. C. Strinati, and S. Barbarossa, "The fog balancing: Load distribution for small cell cloud computing," in *Proc. IEEE 81st Veh. Technol. Conf. (VTC Spring)*, Glasgow, Scotland, May 2015, pp. 1–6.

[49] M. Yemini and A. J. Goldsmith, "'Fog' optimization via virtual cells in cellular network resource allocation," 2019, *arXiv:1901.06669*.

[50] S. Borst, V. Gupta, and A. Walid, "Distributed caching algorithms for content distribution networks," in *Proc. IEEE INFOCOM*, San Diego, CA, USA, Mar. 2010, pp. 1–9.

[51] L. Fleischer, M. X. Goemans, V. S. Mirrokni, and M. Sviridenko, "Tight approximation algorithms for maximum general assignment problems," in *Proc. 17th Annu. ACM-SIAM Symp. Discrete Algorithm*, Miami, Florida, 2006, pp. 611–620.

[52] S. Ioannidis and E. Yeh, "Adaptive caching networks with optimality guarantees," *IEEE/ACM Trans. Netw.*, vol. 26, no. 2, pp. 737–750, Apr. 2018.

[53] D. Malak, F. V. Mutlu, J. Zhang, and E. M. Yeh, "Transmission delay minimization via joint power control and caching in wireless HetNets," in *Proc. 19th Int. Symp. Model. Optim. Mobile, Ad Hoc, Wireless Netw. (WiOpt)*, Oct. 2021, pp. 1–8.

[54] S. Fedor and M. Collier, "On the problem of energy efficiency of multi-hop vs one-hop routing in wireless sensor networks," in *Proc. 21st Int. Conf. Adv. Inf. Netw. Appl. Workshops (AINAW)*, Niagara Falls, ONT, Canada, 2007, pp. 380–385.

[55] U. M. Peşović, J. J. Mohorko, K. Benkiç, and Z. F. Çuçej, "Single-hop vs. multi-hop—Energy efficiency analysis in wireless sensor networks," in *Proc. 18th Telecommun. Forum*, Belgrade, Serbia, Nov. 2010, pp. 1–4.

[56] M. Haenggi and D. Puccinelli, "Routing in ad hoc networks: A case for long hops," *IEEE Commun. Mag.*, vol. 43, no. 10, pp. 93–101, Oct. 2005.

[57] 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, Apr. 2012.

[58] Y. Liu, Y. Li, Q. Ma, S. Ioannidis, and E. Yeh, "Fair caching networks," *Perform. Eval.*, vol. 143, pp. 102–138, Nov. 2020.

[59] T. S. Rappaport, *Wireless Communications: Principles and Practice*, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 1996.

[60] A. Goldsmith, *Wireless Communications*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

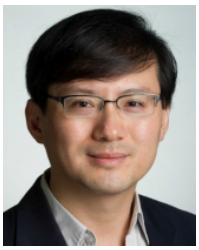
[61] M. Mahloo, P. Monti, J. Chen, and L. Wosinska, "Cost modeling of backhaul for mobile networks," in *Proc. IEEE Int. Conf. Commun. Workshops*, Jun. 2014, pp. 397–402.

[62] 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.
[63] A. Beck, *First-Order Methods in Optimization*, vol. 25. Philadelphia, PA, USA: SIAM, 2017.
[64] A. Ruszczynski, *Nonlinear Optimization*. Princeton, NJ, USA: Princeton Univ. Press, 2011.
[65] D. P. Bertsekas, W. Hager, and O. Mangasarian, *Nonlinear Programming*. Belmont, MA, USA: Athena Scientific, 1998.
[66] A. A. Ageev and M. I. Sviridenko, "Pipage rounding: A new method of constructing algorithms with proven performance guarantee," *J. Combinat. Optim.*, vol. 8, no. 3, pp. 307–328, Sep. 2004.
[67] S. Scheimberg and P. R. Oliveira, "Descent algorithm for a class of convex nondifferentiable functions," *J. Optim. Theory Appl.*, vol. 72, no. 2, pp. 269–297, Feb. 1992.
[68] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2009.
[69] 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.
[70] C. A. Floudas, *Deterministic Global Optimization: Theory, Methods and Applications*, vol. 37. Berlin, Germany: Springer, 2013.
[71] J. Gorski, F. Pfeuffer, and K. Klamroth, "Biconvex sets and optimization with biconvex functions: A survey and extensions," *Math. Methods Oper. Res.*, vol. 66, no. 3, pp. 373–407, Nov. 2007.
[72] C. A. Floudas and V. Visweswaran, "A global optimization algorithm (GOP) for certain classes of nonconvex NLPs—I. Theory," *Comput. Chem. Eng.*, vol. 14, no. 12, pp. 1397–1417, Dec. 1990.
[73] F. H. Clarke, *Optimization and Nonsmooth Analysis*. Philadelphia, PA, USA: SIAM, 1990.
[74] B. T. Polyak, *Introduction to Optimization*, vol. 1. New York, NY, USA: Optimization Software, 1987.
[75] A. Daniilidis and D. Drusvyatskiy, "Pathological subgradient dynamics," *SIAM J. Optim.*, vol. 30, no. 2, pp. 1327–1338, Jan. 2020.
[76] C. Aggarwal, J. L. Wolf, and P. S. Yu, "Caching on the world wide web," *IEEE Trans. Knowl. Data Eng.*, vol. 11, no. 1, pp. 94–107, Jan./Feb. 1999.
[77] J. Wang, "A survey of web caching schemes for the internet," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 29, no. 5, pp. 36–46, Oct. 1999.
[78] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. IT-28, no. 2, pp. 129–137, Mar. 1982.
[79] K. Kamran, E. Yeh, and Q. Ma, "DECO: Joint computation, caching and forwarding in data-centric computing networks," in *Proc. 20th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, Jul. 2019, pp. 111–120.
[80] M. Mahdian, A. Moharrer, S. Ioannidis, and E. Yeh, "Kelly cache networks," in *Proc. IEEE INFOCOM Conf. Comput. Commun.*, Apr. 2019, pp. 217–225.

Derya Malak (Member, IEEE) received the B.S. degree in electrical and electronics engineering (EEE) with a minor in physics from Middle East Technical University in 2010, the M.S. degree in EEE from Koc University in 2013, and the Ph.D. degree in ECE from The University of Texas at Austin in 2017. She has held visiting positions with INRIA and LINCS, Paris, and Northeastern University, and summer internships with Huawei, Plano, TX, USA, and Bell Labs, Murray Hill, NJ, USA. She was a Post-Doctoral Associate with MIT from 2017 to 2019. She was a tenure track Assistant Professor with the Department of ECSE, RPI, from 2019 to 2021. She is currently an Assistant Professor of communication systems with EURECOM, France. Her expertise is in information theory, communication theory, and networking areas. She has developed novel distributed computation solutions, and wireless caching algorithms by capturing the confluence of storage, communication, and computation aspects. She was awarded the Graduate School Fellowship by UT Austin from 2013 to 2017. She was selected to participate in the Rising Stars Workshop for Women in EECS, MIT, in 2018. She received the Best Paper Award from WiOpt 2022 and WiOpt 2023. Her research has been funded by the Huawei Chair Program on Advanced Wireless Systems since 2022, the NSF, the Rensselaer-IBM AI Research Collaboration, and the DARPA Dispersive Computing Programs. She was a recipient of the ERC Starting Grant 2023–2028 on computing nonlinear functions over communication networks (SENSIBILIT).

Faruk Volkan Mutlu received the B.S. degree in electrical and electronics engineering from Middle East Technical University (METU) in 2018 and the M.S. degree in electrical and computer engineering from Northeastern University (NEU) in 2022, where he is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, under the supervision of Prof. Edmund M. Yeh. He also held internship positions with Hewlett Packard Enterprise and Google. As an undergraduate with METU, he contributed to research regarding the emergent age-of-information metric under the supervision of Prof. Elif Uysal. His current expertise is in the fields of information-centric networking and network resource optimization. He has developed novel algorithms for caching and forwarding in networks with diverse cache resources. As a contributor in the NDN for data intensive science experiments (N-DISE) project, he helped build a large scale high-throughput data delivery system for experiments, such as CMS at LHC. He acted as the Web Chair of SIGMETRICS 2020.

Jinkun Zhang received the B.S. degree in microelectronics from Fudan University, China, in 2017. He is currently pursuing the Ph.D. degree in electrical and computer engineering with Northeastern University, USA, under the supervision of Prof. Edmund M. Yeh. He led several research projects in jointly optimizing network forwarding, caching and computation placement strategies, as well as managing wireless power allocation. He is also interested in network game-theoretical behavior and pricing. His research interests include information-centric networking (ICN), network resource optimization, dispersed computing, and distributed machine learning.



Edmund M. Yeh (Senior Member, IEEE) received the B.S. degree (Hons.) and Phi Beta Kappa in electrical engineering from Stanford University in 1994, the M.Phil. degree in engineering from Cambridge University on the Winston Churchill Scholarship in 1995, and the Ph.D. degree in electrical engineering and computer science from MIT under Prof. Robert Gallager in 2001. He was an Assistant Professor and an Associate Professor of electrical engineering, computer science, and statistics with Yale University. He is currently a Professor of electrical and computer engineering with Northeastern University, with a courtesy appointment with the Khoury College of Computer Sciences. He was a recipient of the Alexander von Humboldt Research Fellowship and the Army Research Office Young Investigator Award. He has received four best paper awards, such as WiOpt 2023, ACM ICN 2017, IEEE ICC 2015, and IEEE ICUFN 2012. He served as the TPC Co-Chair for ACM MobiHoc 2021 and the General Chair for ACM SIGMETRICS 2020. He has served as an Area Editor for IEEE TRANSACTIONS ON INFORMATION THEORY and an Associate Editor for IEEE/ACM TRANSACTIONS ON NETWORKING, IEEE TRANSACTIONS ON MOBILE COMPUTING, and IEEE TRANSACTIONS ON NETWORK SCIENCE AND ENGINEERING. He has served as a Treasurer and the Secretary of the Board of Governors of the IEEE Information Theory Society. He is an IEEE Communications Society Distinguished Lecturer.