Joint Routing and Resource Allocation for Millimeter Wave Picocellular Backhaul

Maryam Eslami Rasekh, *Member, IEEE*, Dongning Guo, *Senior Member, IEEE*, and Upamanyu Madhow, *Fellow, IEEE*

***Abstract*— Picocellular architectures are essential for providing the spatial reuse required to satisfy the ever-increasing demand for mobile data. A key deployment challenge is to provide backhaul connections with sufficiently high data rate. Providing wired support (e.g., using optical fiber) to pico base stations deployed opportunistically on lampposts and rooftops is impractical, hence wireless backhaul becomes an attractive approach. A multihop mesh network comprised of directional millimeter (mm) wave links is considered here for this purpose. Such networks are well suited for scaling backhaul data rates due to the abundance of spectrum in the mm wave bands, and the ability to form highly directional, electronically steerable beams. The backhaul design problem is formulated as one of joint routing and resource allocation, accounting for mutual interference across simultaneously active links. A computationally tractable formulation is developed by leveraging the localized nature of interference and the provable existence of a sparse optimal allocation. Numerical results are provided for topologies modeling urban and suburban settings.**

***Index Terms*— Millimeter wave, 5G, wireless backhaul, mesh network, resource allocation, medium access control, routing, interference, mixed-integer programming.**

I. Introduction

THE wireless industry is striving to keep up with mobile data demand from smart devices and data-hungry applications. Picocellular architectures comprised of closely-spaced access points with intense spatial reuse play a critical role in the evolution of mobile systems, particularly in high-density urban and suburban environments [1]. LTE data rates are projected to approach Gigabits per second peak rates through carrier aggregation, which is likely to be extended further in next generation networks by using 60 GHz - or other unlicensed millimeter (mm) wave bands - directly from pico base station to the mobile [2], assuming that significant challenges due to blockage and mobility can be overcome. Indeed, a recent interference analysis for such networks [3] indicates that capacities of the order of terabits per second per kilometer (along a single urban canyon) can be obtained with only a few GHz of spectrum by taking advantage of the aggressive spatial reuse enabled by highly directional mm wave links. Moreover, this capacity roughly adds up across parallel canyons, given the strong isolation provided by building blockage.

Delivering such high data rates to mobile users requires that the pico base stations have a sufficiently high-capacity backhaul connection to the Internet. For opportunistic picocellular deployments on lampposts and rooftops, providing optical fiber connectivity for each base station is practically impossible in the foreseeable future, so that wireless backhaul becomes a natural choice [4]–[6].

In this paper, we consider a mesh network with highly directive mm wave links as a means of extending backhaul from wired gateways to the picocell access points. Each access point is a node in the wireless mesh network and is connected to neighboring nodes through high-speed directional mm wave links. The objective is to route traffic between base stations and gateways through multihop paths, such that each picocell can support a given level of downlink and uplink throughput on the access links to mobile devices. We assume that directional antennas are used on each backhaul link, but because of half duplex communication and residual interference, these links cannot be treated as wires. In particular, in long urban canyons where even non-contiguous links are likely to be aligned, the interference between distant links must be taken into account in deriving the optimal resource allocation and routing.

A. Approach and Contributions

We formulate the wireless mesh backhaul design problem as a joint routing and resource allocation optimization, with the goal of maximizing the access rate at the base stations, while accounting for the mutual interference between simultaneously active links. Our framework applies to any mesh backhaul network via the following abstraction: For any set of simultaneously active links, we must be able to compute the achievable data rate on each active link while accounting for the interference from other active links. However, the computational complexity depends on the nature of the interference patterns, which depends on the propagation environment, the antenna patterns, and the carrier frequency. Specifically, the highly directive nature of mm wave links leads to localized and sparse interference that we are able to exploit for efficient computation of the optimal solution. We illustrate

Manuscript received September 13, 2018; revised May 10, 2019 and October 2, 2019; accepted October 10, 2019. Date of publication October 29, 2019; date of current version February 11, 2020. This work was supported in part by the National Science Foundation under Grant CCF-1423040, Grant CNS-1317153, and Grant CNS-1518812, and gifts from Facebook, Qualcomm, and Futurewei Technologies. The associate editor coordinating the review of this article and approving it for publication was S. Kompella. *(Corresponding author: Maryam Eslami Rasekh.)*

M. Eslami Rasekh and U. Madhow are with the Department of Electrical and Computer Engineering, University of California Santa Barbara, Santa Barbara, CA 93106 USA (e-mail: rasekh@ucsb.edu; madhow@ece.ucsb.edu).

D. Guo is with the Department of Electrical and Computer Engineering, Northwestern University, Evanston, IL 60208 USA (e-mail: dguo@northwestern.edu).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TWC.2019.2948624

our approach for two settings, urban and suburban, which exhibit interference patterns that are quite different. In built-up urban environments, the highly directive nature of the links, and the ease of blockage of mm waves by buildings, imply that each link incurs interference only with links within the same street. However, the interference among such links can be significant due to their similar alignment. In more open suburban environments, on the other hand, a larger number of links can cause interference on each other, but links are less likely to be aligned, so that interference is typically severely attenuated by antenna directionality at the transmitter and the receiver.

Key novel contributions of the proposed architecture and analysis include the following:

- *Accurate interference modeling:* Environment-specific propagation models are used to derive a realistic interference graph for the network and the optimal level of spatial reuse is maintained, yielding the highest possible backhaul throughput upon deployment. The solutions proposed here can be easily generalized to a variety of network structures with different antenna patterns and arbitrary interference models.
- *Arbitrary (fixed-rate) traffic flow:* We first develop a framework for *downlink* traffic routing and resource allocation. We then extend the formulation to perform joint uplink and downlink optimization. Additional flows between nodes in the network can be included in the optimization by introducing a new commodity for each flow. While we only consider fixed-rate traffic flows, extension to any concave objective of the rate vector is possible in the proposed framework.
- *Arbitrary topologies:* The proposed optimization framework applies to arbitrary network structures with limited node degrees. The networks in our simulations are comprised of several gateways that are connected to all nodes through multihop paths. Each access point is connected to all neighbors that are close enough to form a connection (the number of neighbors is not large in practice). The existence of various paths from each pico base station to different gateways provides redundancy in network resources and improves backhaul reliability, allowing the network to adapt to disruption of links due to blockage or hardware failure. Backhaul in future cellular networks may constitute a mix of wireless *and wired* links. Wired links can easily be incorporated in the proposed framework as links with capacity constraint that neither cause nor are affected by interference.
- *Versatile and efficient optimization framework:* The problem of resource allocation for optimizing backhaul throughput is first formulated as a linear program, the dimensions of which grow exponentially with network size. Demonstrating the existence of a sparse solution to this linear program, we formulate an equivalent mixed-integer linear program that scales linearly with network size. The proposed formulation is able to solve relatively large networks with hundreds of nodes in a short period of time.

B. Related Work

Multihop wireless mesh networks have received extensive attention in the literature, typically with sensible heuristics for solving the associated joint routing and scheduling problems. Non-contention-based medium access protocols that rely on link scheduling and resource allocation have been studied for decades. Various works such as [7], [8] consider the link scheduling problem independently assuming routing is known beforehand, while others attempt to jointly optimize scheduling and routing. In modeling the interference behavior of links, the vast majority of these studies assume a binary effect, commonly known as the "protocol model" as described in [9]: Any two links either collide completely (mostly by contradicting the half duplex constraint or, in omnidirectional settings, falling within some interference radius) and have to be orthogonalized in resources, or have zero interference [10]–[19]. In [14], [15] a transmission is assumed to fail if the number of active transmitters within interfering range of its receiver is above a threshold, which is a modified version of the collision model.

Few papers take note of the residual interference between links in the network and allow for capacity-SINR trade-off in their allocation optimization; this can lead to suboptimal solutions and lost throughput as demonstrated in [20]. In [21], confining the analysis to the low SINR regime, link throughput is approximated as a linear function of SINR. This approximation is not suitable for the high-speed point-to-point links employed for wireless backhauling.

Several papers consider the specific problem of multihop mesh backhauling with directional antennas, but limit the number of links on each node to a single steerable beam [4], [10], [14], [15], [17], [22]. Other works such as [12], [19] allow simultaneous multi-neighbor communication but designate a fixed discrete partitioning of resources, generally bandwidth, and limit each link to a single partition at any time. Coarse discretization of resources and single channel confinement of link transmissions both prevent achieving the true capacity of the network, and orthogonalizing same-node links severely limits the potential for spatial reuse, especially as such limitations are not present when employing highly directional antennas that reliably isolate many simultaneous same-node transmissions. In contrast, our work allows continuous partitioning and places no constraint on which part(s) of the resources each node may use other than what is imposed by the physical interference model, thereby realizing the true capacity of the network and facilitating the spatial reuse required for backhaul delivery.

Increasing cellular capacity through self-backhauled small cells has been the subject of many previous works. Some studies consider placement of relay nodes inside cells to improve signal quality at cell edges [13], [23], [24], yet the capacity boost obtained through simple radio frequency (RF) amplification and relaying is limited. Other studies consider addition of small-cell base stations inside a macrocell, each receiving wireless backhaul directly from the wired macro-BS, forming a backhaul network with star topology [4], [25], [26]. In [27], a multihop network is considered in the form of a row of nodes inside one street along with one wired node

that acts as a gateway. In [13], the network comprises one base station, relay nodes which do not generate data, and end nodes which do not relay data. These simplified structures are useful to provide insight into the capacity of wirelessly-backhauled picocells, but evaluating and optimizing a realistic network requires increasing the scope to include multihop paths along different streets and networks with several gateways.

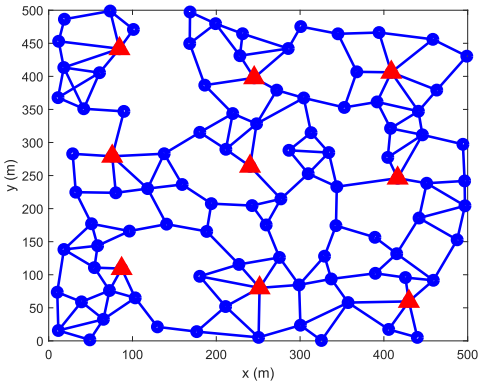
In contrast to the heuristics employed in prior literature on multihop mesh networks, we are able to compute the optimal solution by considering all possible link activation patterns. The idea of allocating resources to "cliques" of links has been around for decades [28], yet it has also been shown that finding the true optimal allocation and routing is an NP-hard problem, the complexity of which scales exponentially with network size [29], [30]. Several studies tackle the intractable problem of interference-aware routing and resource allocation using various heuristics [10], [30]–[36]. These include either relaxing the interference model to a conflict graph and solving the resulting problem using edge coloring algorithms, or column generation based methods that iteratively search for "good" patterns and solve the allocation problem over the discovered patterns in each iteration. These approaches often fail to obtain the exact optimal throughput due to their limitations in modeling or greedy procedure.

Finally, in [37], a mm wave multihop backhaul network was modeled as a uniform square grid of nodes to provide analytical insight into throughput capacity as a function of the size of the cluster supported by each gateway. An interesting observation was that backhaul capacity is not diminished as a result of residual interference, and, with careful scheduling of interfering links, the interference-free capacity of the network can be achieved. We demonstrate the importance of interference avoidance in scheduling and note that the proposed framework provides such a schedule for their specific topology.

Inspired by the approach used in [38]–[41] for optimizing downlink spectrum allocation in cellular networks, our starting point is the problem of resource allocation among all possible subsets of links. Then, following [42], we reformulate the convex combinatorial resource allocation problem into a scalable mixed integer optimization problem. The problem considered here is fundamentally different from that of [38]–[42] due to the multihop nature of the network and the added problem of routing both uplink and downlink traffic.

The present paper consolidates and significantly extends the work in our prior conference publication [43], where we present a combinatorial optimization framework for optimal resource allocation and routing. In [43], resources are divided between all possible activation patterns, or *subsets* of simultaneously active links. This blows up the problem size exponentially, going from a scheduling of L links to 2^L possible link combinations, but enables formulation of backhaul scheduling as a linear program. Solving this problem is straightforward for small networks but quickly becomes intractable for larger graphs, and requires suboptimal clustering of the larger network. One important result, guaranteed by Caratheodory's theorem [44], is that a sparse solution exists for the combinatorial formulation with at most N active patterns, where N is the number of nodes in the network. In this paper,

(a)



(b)

Fig. 1. Example of (a) a 48-node, 4-gateway portion of an imagined picocellular backhaul network in Manhattan, (b) a randomly generated suburban network of 100 nodes and 9 gateways in a 500 m×500 m area. Gateways are identified by triangles and non-gateway nodes by circles.

we first revisit the problem in [43] and provide a scalable optimization framework which exploits the existence of an N-sparse solution. We reformulate the problem in terms of *local* interference parameters. The result is a binary linear program (BLP) that scales near-linearly with network size, and can be solved relatively quickly using branch and bound techniques for larger networks with multiple gateways and hundreds of nodes. We also incorporate joint scheduling for uplink and downlink traffic, while the analysis of [43] only optimizes for downlink support.

C. Outline

The remainder of this paper is organized as follows. The system model is described in Section II. In Section III we formulate the routing and scheduling problem and prove the existence of a sparse solution. We then describe the construction of a scalable formulation on this basis in Section IV. We present and discuss simulation results in Section V. Conclusions and possible directions for future work are discussed in Section VI.

II. System Model

We envision an outdoor picocellular network with cell radii as small as tens of meters. The pico base stations are placed opportunistically on existing structures such as lampposts, building walls, and ceilings. Each base station, or "node", is connected to several nearby nodes (neighbors) through directional mm-wave links. Only a fraction of nodes are

equipped with wired backhaul connection to the communication infrastructure; these nodes operate as gateways. The downlink and uplink data of each node is routed through a multihop path in the mesh to one or more of the gateway nodes. Examples of this structure are shown in Fig. 1 where the gateways are marked by triangles and non-wired base stations by circles. Each line connecting two neighboring nodes represents two wireless links, one in each direction. Phased array antennas are utilized at either end of the link to maintain directional transmit and receive radiation patterns and reduce interference. We define the nominal link SNR and nominal link rate as the SNR and throughput of links in the absence of interference, and transmit power is controlled such that these values are the same for all links.

We denote by $\Gamma = \{1, \ldots, N\}$ the index set of all N non-gateway nodes in the network and by $\Lambda = \{1, \ldots, L\}$ the index set of all links in the network. As far as interference allows, any number of links connected to a node can be active simultaneously. However, because a transmitted signal is generally strong enough to saturate all co-located receivers, communication is half-duplex, namely, the links connected to one node can transmit simultaneously, or receive simultaneously, but at no time do some links transmit and some receive. As the opposing directions between two nodes are regarded as two separate links in the proposed optimization framework, the duplexing constraints can be easily incorporated by assuming infinite (or disabling) interference levels for such conflicting links.

In order to find the optimal allocation for interference management and routing, we apply a combinatorial approach where the available resources (e.g., time and/or frequency) are divided between different *subsets* of links with the objective of maximizing the minimum backhaul throughput delivered to all nodes in the network. A general scheduling framework can be defined as follows. The total time in each frame is divided into fragments of variable lengths and during each fragment (or slot) a certain combination of links are active. We denote each possible partitioning of links into on and off by an "activation pattern", which determines the amount of interference each link is subject to, and hence its data rate. The resource management problem is equivalent to allocating to every possible activation pattern an appropriate fraction of all resources (which may be none). For every possible activation pattern $A \subset \Lambda$, let $x_A \in [0, 1]$ denote the fraction of resources allocated to that pattern, which must satisfy the resource constraint,

$$\sum_{A \subset \Lambda} x_A = 1.$$

The problem of allocating resources to L links becomes that of allocating non-overlapping portions to the $2^L - 1$ non-empty subsets of links.

While the uplink or downlink rates on access links for a given *mobile* user can be bursty, and vary significantly across small timescales, the backhaul capacity demand at each base station, generated by aggregating its access link traffic, is smoother. Thus, our framework supports a constant demand at each base station over each optimization period, while allowing the flexibility of allocating resources differently across different base stations. Specifically, as we shall see, we may specify that base station i receive backhaul capacity allocation $\alpha_i d$, where d is a quantity to be maximized by joint routing and resource allocation.

We consider two distinct deployment scenarios, modeling suburban and urban settings. In order to quantify link throughput, we need to determine the spectral efficiency of a link for a given activation pattern. Approximating the sum of the interference plus noise as Gaussian, the spectral efficiency of link l depends only on its signal-to-interference-plus-noise-ratio (SINR), and is given by

$$\text{SINR}_l = \frac{S_l}{\sigma^2 + \sum_{k:\text{active links}} I_{k \to l}}.$$

Here S_l and σ^2 are the signal and noise PSD respectively, and $I_{k \to l}$ denotes the interference caused by link k on link l. Without loss of generality, we normalize noise power to unity ($\sigma^2 = 1$) and determine signal and interference power in proportion to noise. For both urban and suburban scenarios, we assume links have nominal (interference-free) SNR of SNR_{nom} (set to 10 dB in our numerical results), and that each antenna array has N omnidirectional elements with half-wavelength spacing ($N = 32$ in our numerical results). Thus the signal level of all links is

$$S_l = \text{SNR}_{\text{nom}},$$

and the transmit power of link l is calculated by

$$P_l = S_l (G_T G_R \frac{\lambda}{4\pi D_l})^{-2}$$

where D_l is the distance of link l, $G_T = G_R = N$ are the transmit and receive antenna gains, and λ is the carrier wavelength. As all parameters except D_l are the same for all links, we can simplify interference derivations by noting that

$$\frac{P_l}{P_k} = \frac{D_l^2}{D_k^2}.$$

The normalized interference caused by link k in link l is derived from the channel model as follows.

Suburban channel model. In the suburban setting, antennas are placed on rooftops that are of relatively similar heights. As a result, the street geometry does not have a significant effect on the channel as links are assumed to be line-of-sight (LOS) without any reflective structures between antennas, as shown in Fig. 2. The free-space propagation model is thus used to model the channel between different nodes and the only factors determining signal and interference strength are the radiation pattern of antennas and distances between nodes, allowing us to describe interference of link k on link l by

$$\begin{aligned} I_{k \to l} &= P_k \, G_T(\theta^{\text{dep}}_{k \to l}) \, G_R(\theta^{\text{arr}}_{k \to l}) \left(\frac{\lambda}{4\pi D_{k \to l}} \right)^2 \\ &= S_l \, G_T(\theta^{\text{dep}}_{k \to l}) \, G_R(\theta^{\text{arr}}_{k \to l}) \left(\frac{D_k}{D_l D_{k \to l}} \right)^2 \\ &= \text{SNR}_{\text{nom}} \, G_T(\theta^{\text{dep}}_{k \to l}) \, G_R(\theta^{\text{arr}}_{k \to l}) \left(\frac{D_k}{D_l D_{k \to l}} \right)^2 \end{aligned}$$

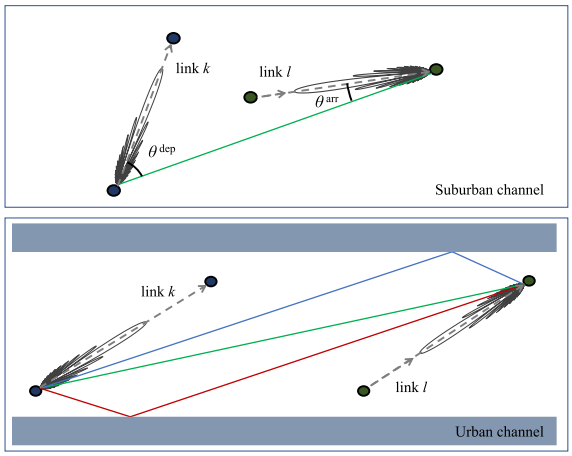


Fig. 2. Ray tracing channel model for urban and suburban scenarios.

where $D_{k\to l}$ is the line-of-sight distance between the transmitter of link k and the receiver of link l, and $\theta^{\text{dep}}_{k\to l}$ and $\theta^{\text{arr}}_{k\to l}$ denote the angles of departure and arrival of the interference path relative to the pointing direction of transmit and receive antennas, respectively.

Urban channel model. In the high-rise urban setting, antennas are mounted on below-rooftop-level structures such as lamp-posts, traffic lights, and building walls. In this case, a street-canyon model is considered, modeling the channel as a combination of the LOS path and single-bounce reflections from the two canyon walls and ground, as shown in Fig. 2. Transmit and receive radiation patterns and Fresnel (specular) reflection loss are accounted for in this model, and the total interference power is the summation of the power conveyed from the four paths,

$$I_{k\to l} = \text{SNR}_{\text{nom}} \frac{D_k^2}{D_l^2} \sum_{i=0}^{3} \gamma_{\text{ref}}^i G_T(\theta^{i,\text{dep}}_{k\to l})\, G_R(\theta^{i,\text{arr}}_{k\to l})\, \frac{1}{(D^i_{k\to l})^2}.$$

In this equation, γ^i_{ref} is the reflection loss of path i (equal to 1 for the direct path), determined by its angle of incidence and the dielectric constant of the street walls and ground (set to 6.5 in our simulations). The departure and arrival angles, and angle of incidence for reflections are dictated by the network geometry. Urban streets are assumed to be 25 m wide and the distance of each node from street wall and ground is chosen uniformly over $(4, 21)$ m and $(5, 8)$ m respectively.

In simulations, we consider the topology of Fig. 1(a) for the urban scenario, and randomly generated topologies similar to Fig. 1(b) for the suburban scenario. We use a heuristic algorithm to generate random networks for testing our framework, tuning parameters with trial and error until plausible geometries are achieved. In this procedure, nodes are placed uniformly at random in a square area with a minimum link distance of 20 m. For each non-gateway node, a random number is drawn between 3 and 5, then the node is connected to that number of its nearest neighbors with bidirectional links, and links that are longer than a threshold are discarded to limit the variation of link distances. If two links on a node are too close in the angular domain (within two beamwidths), the longer link is dropped. The actual degree of a node may be larger than the number of neighbors it connects to, since additional links may be established by its neighbors. Each bidirectional link is actually two separate links individually indexed in the set Λ. The maximum link length threshold is tuned to obtain a well connected graph and is set high enough to ensure that no node is left disconnected from the network. Gateway nodes create connections with at most 6 closest neighbors whose distance is below the link establishment threshold. This is an appropriate design practice to increase the backhaul capacity of the network, as gateway links are traffic bottlenecks.

The number of gateways is chosen to be a fraction η of the total number of nodes, and gateway nodes are defined by choosing the closest nodes to uniformly placed anchor points in the area. While this generating scheme obtains relatively realistic topologies, the occurrence of sub-par configurations that impose bottlenecks to service of some areas is possible, which can significantly decrease the max-min capacity of the network. In practice, node placements would be optimized to some extent to improve the capacity of the network and better performance than the simulated outcomes can be expected.

III. Combinatorial Formulation of the Allocation Problem

In this section we formulate an optimization problem that maximizes the minimum backhaul throughput delivered to every node subject to interference and resource constraints.

When pattern $A \subset \Lambda$ is active, the throughput of link $l \in A$ is equal to,

$$\gamma_{l,A} = \log\left(1 + \frac{S_l}{\sigma^2 + \sum_{k\in A\setminus\{l\}} I_{k\to l}}\right). \tag{1}$$

While $\gamma_{l,A}$ is left undefined if $l \notin A$ to avoid redundant parameters, it is equivalent (and natural) to think of it as equal to 0 if $l \notin A$. The total data rate of link l, i.e., data transferred in a unit of time, is therefore obtained by,

$$r_l = \sum_{A\subset\Lambda: l\in A} x_A \gamma_{l,A}.$$

A. Optimization of Downlink Only

In the downlink, each node is considered as a sink of its own cell's traffic. The gateway nodes are sources with no throughput constraint. Although this is a network of multiple "commodities", each intended for one non-gateway node and potentially delivered via a combination of multiple flows, it is equivalent to a network of a single commodity by using the following insight: We add a virtual source node representing the core (wired) network where all flows originate, with a link of unlimited capacity to each of the gateway nodes. Since this is the sole source node in the network, *any* data that flows into *any* node has invariably originated in this virtual node and is therefore downlink backhaul traffic. The reader is referred to [45] for a detailed discussion of network commodities. The downlink data rate delivered to node i, normalized to system bandwidth, is thus equal to

$$d_i = \sum_{l\in I_i} r_l - \sum_{k\in O_i} r_k, \quad i \in \Gamma \tag{2}$$

where I_i is the set of links that flow into[1] node i and O_i is the set of links that flow out of it.

If allocations were in the frequency domain, the spectral efficiency parameters $\gamma_{l,A}$ would depend on the allocation variables x_A as the noise power may depend on the bandwidth. Time domain allocation thus simplifies the formulation.

The network utility is in general a function of the rate vector $[r_1, \ldots, r_L]^T$. Our formulation provides the flexibility to provide different backhaul throughput to different nodes by guaranteeing throughput $\alpha_i d$ to node i, with optimal allocation maximizing d. This allocation can be obtained by solving the following optimization problem:

$$\begin{aligned}
\underset{[x_A],[r_l],[d_i],d}{\text{maximize}} \quad & d && \text{(3a)}\\
\text{subject to} \quad & \sum_{A\subset\Lambda} x_A = 1, && \text{(3b)}\\
& r_l = \sum_{A\subset\Lambda: l\in A} x_A \gamma_{l,A}, \quad l \in \Lambda && \text{(3c)}\\
& d_i = \sum_{l\in I_i} r_l - \sum_{k\in O_i} r_k, \quad i \in \Gamma && \text{(3d)}\\
& d_i \geq \alpha_i d, \quad i \in \Gamma && \text{(3e)}\\
& x_A \geq 0, \quad A \subset \Lambda && \text{(3f)}
\end{aligned}$$

where r_l is the data rate on link l. The weighting factor α_i can be used to provide larger backhaul throughput for high-traffic hotspots if needed. (In our simulations, we assume uniform traffic at all nodes and set $\alpha_i = 1, \forall i$.) Note that the node flow constraints only apply to non-gateway nodes. Relaxing the resource, rate, and throughput equalities, (3b), (3c), and (3d) to inequality $\leq$ is inconsequential.

Both the objective and constraints in the optimization are linear, therefore an optimal allocation can be found by solving a linear program with around 2^L variables. Of course, if all of these possible activation patterns receive non-zero time allocation, the complexity of scheduling would grow exponentially with network size. However, there is always a sparse optimal solution to (3) in the following sense:

Theorem 1: For a network of L links, N non-gateway nodes, and one or more gateways, there exists a solution $[x_A]_{A\subset\Lambda}$ to the downlink scheduling problem of (3) that is at most N-sparse, i.e., $x_A = 0$ for all but (at most) N activation patterns A.

Proof: Consider the N dimensional vector $\mathbf{d} = [d_1, \ldots, d_N]^T$ that is a feasible point for (3) achieved by the allocation $[x_A]$. For the entries of this vector, we have,

$$\begin{aligned}
d_i &= \sum_{l\in I_i} r_l - \sum_{k\in O_i} r_k\\
&= \sum_{l\in I_i} \sum_{A\subset\Lambda: l\in A} x_A \gamma_{l,A} - \sum_{k\in I_i} \sum_{A\subset\Lambda: k\in A} x_A \gamma_{k,A}\\
&= \sum_{A\subset\Lambda} x_A \left(\sum_{l\in I_i\cap A} \gamma_{l,A} - \sum_{k\in O_i\cap A} \gamma_{k,A} \right).
\end{aligned}$$

[1]Not to be confused with $I_{k\to l}$ which denotes the interference of link k on link l.

The maximum utility is determined by $\mathbf{d}$, which can be written as a convex combination of $2^L - 1$ vectors of dimension N as,

$$\begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_N \end{bmatrix} = \sum_{A\subset\Lambda} x_A \begin{bmatrix} \sum_{l\in I_1\cap A} \gamma_{l,A} - \sum_{k\in O_1\cap A} \gamma_{k,A} \\ \sum_{l\in I_2\cap A} \gamma_{l,A} - \sum_{k\in O_2\cap A} \gamma_{k,A} \\ \vdots \\ \sum_{l\in I_N\cap A} \gamma_{l,A} - \sum_{k\in O_N\cap A} \gamma_{k,A} \end{bmatrix}.$$

Hence $\mathbf{d}$ resides in a polyhedron in $\mathbb{R}^N$ with up to $2^L - 1$ corner points. Noting the resource constraint, $\sum_{A\subset\Lambda} x_A = 1$, Caratheodory's theorem [44] states that $\mathbf{d}$ can be written as a convex combination of at most $N + 1$ of these $2^L - 1$ points. Therefore there exists an allocation $[x'_A]$ satisfying $\sum_{A\subset\Lambda} x'_A = 1$ with at most $N+1$ nonzero entries that yields the same node service vector as $[x_A]$. An optimal $\mathbf{d}$ cannot be an interior point of the polyhedron; otherwise one may strictly increase all of its elements to a boundary point, which increases the objective d. This implies that there exists an N-sparse optimal allocation, hence the proof of Theorem 1. □

For the mm wave networks considered here, the interference matrix can be very sparse, resulting in many possible optimal allocations when we solve the linear program. By adding very small random fluctuations to the interference matrix, the solution can be made unique with high probability, and since a sparse solution is guaranteed to exist, we find one using this trick.

One problem that arises when solving (3) is that the objective function does not penalize suboptimal routing as long as the delivered throughput is unchanged. As a result, a shorter path toward a node is not differentiated from a longer path with more hops, which can result in unnecessarily long paths and excessive latency and power consumption. This can be prevented by adding a linear term to the objective that implicitly penalizes delay. Assuming transfer of data on each link represents one unit of delay in the network, the sum of all link data rates can be taken as a linear proxy for delay and power consumption. This is done by changing the objective of (3a) to

$$d - \lambda \sum_{l\in\Lambda} r_l, \tag{4}$$

where the weighting factor λ is chosen to be small enough to ensure service rate, d, is always prioritized over the delay penalty term. It is shown in [43] that a sufficient condition to enforce this priority is

$$\lambda < \frac{1}{L \sum_i \alpha_i} \tag{5}$$

which, in the case of uniform service to all nodes ($\alpha_i \equiv 1$), simplifies to

$$\lambda < \frac{1}{LN}.$$

B. Joint Uplink-Downlink Optimization

In this section, we extend the formulation of (3) to include both uplink and downlink. While the single commodity model no longer applies, we can add a second commodity representing uplink data that can originate at any non-gateway node

but terminates at a virtual *sink* node connected to all gateway nodes through infinite-capacity links. For this commodity, all traffic will ultimately end up at the virtual sink (core network) and is therefore uplink data, meaning the difference between outgoing and incoming traffic at any node is the uplink throughput provided for that node. The downlink and uplink commodities must share network resources, therefore we formulate the joint optimization by defining two sets of link rate variables, $[r_l^{\mathrm{d}}]$ and $[r_l^{\mathrm{u}}]$, corresponding to downlink and uplink service rates, $[d_i]$ and $[u_i]$, that satisfy

$$\sum_{l\in I_i} r_l^{\mathrm{d}} - \sum_{k\in O_i} r_k^{\mathrm{d}} = d_i,$$
$$\sum_{l\in O_i} r_l^{\mathrm{u}} - \sum_{k\in I_i} r_k^{\mathrm{u}} = u_i.$$

The rate on each link is the sum of the rate supporting downlink and uplink data, and (3c) is modified to

$$r_l^{\mathrm{d}} + r_l^{\mathrm{u}} = \sum_{A\subset\Lambda: l\in A} x_A \gamma_{l,A}.$$

The optimal allocation is hence obtained from solving the following optimization problem, wherein the constants α_i and β_i determine the relative downlink and uplink traffic of different nodes.

$$\underset{[x_A],[r_l^{\mathrm{d}}],[r_l^{\mathrm{u}}],[d_i],[u_i],c}{\text{maximize}} \quad c \tag{6a}$$
$$\text{subject to} \quad \sum_{A\subset\Lambda} x_A \le 1 \tag{6b}$$
$$r_l^{\mathrm{d}} + r_l^{\mathrm{u}} = \sum_{A\subset\Lambda: l\in A} x_A\gamma_{l,A}, \quad l\in\Lambda \tag{6c}$$
$$\sum_{l\in I_i} r_l^{\mathrm{d}} - \sum_{k\in O_i} r_k^{\mathrm{d}} = d_i, \quad i\in\Gamma \tag{6d}$$
$$\sum_{l\in O_i} r_l^{\mathrm{u}} - \sum_{k\in I_i} r_k^{\mathrm{u}} = u_i, \quad i\in\Gamma \tag{6e}$$
$$d_i \ge \alpha_i c, \quad u_i \ge \beta_i c, \quad i\in\Gamma \tag{6f}$$
$$r_l^d \ge 0, \qquad r_l^u \ge 0.\ l\in\Lambda \tag{6g}$$

For this formulation, Caratheodory's theorem cannot be applied to the node rate inequalities as easily as in Theorem 1. However, a similar argument can be made for the link rate vector $[r_1^{\mathrm{d}} + r_1^{\mathrm{u}}, \ldots, r_L^{\mathrm{d}} + r_L^{\mathrm{u}}]$ that is a convex combination of the $2^L - 1$ points (enumerated by A) in L dimensional space, $[\gamma_{1,A}, \ldots, \gamma_{L,A}]^T$. For any allocation $\mathbf{x} = [x_A]_{A\subset\Lambda}$ that is able to support rate vectors $\mathbf{r}^{\mathrm{d}}$ and $\mathbf{r}^{\mathrm{u}}$ and, equivalently, node downlink and uplink vectors $\mathbf{d}$ and $\mathbf{u}$, there exists an L-sparse allocation $\mathbf{x}'$ that provides the exact same link rates and uplink and downlink node service rates. Thus a solution with no more than L active patterns can be guaranteed to exist for any target rate vector pair.

While it is not of interest to the backhaul scenario, any flow between some source node and some sink node can similarly be incorporated in the formulation by introducing a new commodity with its own rate variables, $\{r_l^{\mathrm{f}}\}$, and flow constraints for the sink, source, and other nodes as follows.

$$\sum_{l\in O_{\text{source}}} r_l^{\mathrm{f}} - \sum_{k\in I_{\text{source}}} r_k^{\mathrm{f}} = c_f,$$
$$\sum_{l\in O_{\text{sink}}} r_l^{\mathrm{f}} - \sum_{k\in I_{\text{sink}}} r_k^{\mathrm{f}} = -c_f,$$
$$\sum_{l\in O_i} r_l^{\mathrm{f}} - \sum_{k\in I_i} r_k^{\mathrm{f}} = 0, \quad i\in\Gamma\backslash\{\text{source},\text{sink}\}.$$

The rate variables would then be added to the left-hand side of (6c) to enforce sharing of network resources between flows.

The formulations presented in this section are suitable for relatively small networks. As L increases to even moderate sized values, the number of variables in the problem grows exponentially until the time or space (memory) complexity becomes unmanageable. In the next section, a scalable reformulation is developed.

IV. A Scalable Reformulation

One characteristic of the network that can be leveraged to reduce the problem size is the localized nature of interference, which allows decoupling of constraints between distant areas of the network. We first define the neighborhood of link l as the set of links, Λ_l, that cause non-negligible interference on it, i.e., whose signal strength at the receiving end of l is above a threshold. Subsequently, the "local activation patterns" of link l are all possible subsets of its neighborhood, $B\subset\Lambda_l$. A local activation pattern B is in effect when all links in B are active *and* all links in $\Lambda_l\backslash B$ are inactive. By this definition, the spectral efficiency of a link only depends on the activation pattern of links in its neighborhood, or its *local* activation pattern. To maintain a pessimistic estimate of throughput, the interference of all links outside the neighborhood are added to the noise and interference level when calculating throughput. In reality, many of the links outside the neighborhood will not be active, but we can ensure this worst-case assumption does not affect the result significantly by setting the threshold to be low enough. There is thus a trade-off between computational complexity and accuracy: choosing a low threshold yields more exact results at the expense of increasing neighborhood size and enlarging the problem. In our simulations, we set the interference threshold to 3 dB below the noise level, assuming a nominal SNR of 10 dB. We find the disparity between the pessimistic and actual throughput to be less than 1% in all simulated cases.

Similar to (3), we define the *local* allocation variable x_l^B that is the resource allocated to local activation pattern $B\subset\Lambda_l$ of link l. When enumerating local activation patterns, the empty set is also counted since a nonempty global pattern may activate none of the links in one neighborhood. The data rate on link l is thus equal to

$$r_l = \sum_{B\subset\Lambda_l: l\in B} x_l^B \gamma_l^B, \quad l\in\Lambda$$

where γ_l^B is the spectral efficiency of link l under local activation pattern B, derived (pessimistically) as

$$\gamma_l^B = \log\left(1 + \frac{S_l}{n + \sum_{k\in B\backslash\{l\}} I_{k\to l} + \sum_{j\notin\Lambda_l} I_{j\to l}}\right).$$

Recall that Theorem 1 ensures the existence of an optimal solution to (3) that activates at most N patterns. We therefore consider a segmentation of the unit time frame to N slots, indexed by $M = \{1, \ldots, N\}$. The global slots are of variable lengths, denoted by $\{y_m\}_{m \in M}$, that satisfy the resource constrains, $\sum_{m \in M} y_m = 1$. Let P_m denote the *global* pattern activated in the m-th slot. Then $P_m \cap \Lambda_l$ is the corresponding *local* pattern in the neighborhood of link l in the m-th slot. These local patterns may overlap and their union is P_m. We define the augmented local allocation variables $x_l^{B,m}$ as the resource allocated to local pattern B of link l in slot m. Evidently, for every $l \in \Lambda$, $m \in M$, and $B \subset \Lambda_l$,

$$x_l^{B,m} = \begin{cases} y_m & \text{if } B = P_m \cap \Lambda_l, \\ 0 & \text{otherwise.} \end{cases}$$

For the throughput calculations to hold, local patterns included in a global pattern must be consistent with each other. To enforce this constraint, we introduce a discrete activation parameter for each local pattern, denoted by $q_l^{B,m}$, which is a *binary* variable that takes the value of 1 when its corresponding local pattern is active in global pattern P_m and 0 otherwise. Activation is enforced by the inequality,

$$x_l^{B,m} \leq q_l^{B,m},$$

and consistency is enforced by limiting the sum of *incompatible* local patterns to 1, allowing at most one of them to be non-zero. The local patterns $B \subset \Lambda_l$ and $A \subset \Lambda_k$, corresponding to activation variables $q_l^{B,m}$ and $q_k^{A,m}$, are "compatible" if and only if,

$$A \cap \Lambda_l = B \cap \Lambda_k,$$

which means any active link in B that happens to be in the neighborhood of link k is also active in A, i.e., the two patterns do not impose contradictory activations on any links in the overlap of their neighborhoods. An example of consistent local patterns is shown in Fig. 3 with neighborhoods depicted as sets and links as elements of these sets. Assuming gray squares are inactive links, local pattern A in the neighborhood of link k is consistent with pattern B_1 in the neighborhood of link l but inconsistent with pattern B_2 in the same neighborhood, since link l_0 is inactive in A but active in B_2.

Using the binary activation parameters defined above, consistency of local allocations can be enforced by the inequality,

$$q_l^{B,m} + q_k^{A,m} \leq 1, \quad \forall A \cap \Lambda_l \neq B \cap \Lambda_k. \tag{7}$$

Thus, the optimization problem of (3) can be reformulated as:

$$\underset{[x_l^{B,m}],[q_l^{B,m}],[y_m],[r_l],d}{\text{maximize}} \quad d \tag{8a}$$

$$\text{subject to } r_l \leq \sum_{m \in M} \sum_{B \subset \Lambda_l : l \in B} x_l^{B,m} \gamma_l^B, \quad l \in \Lambda \tag{8b}$$

$$\alpha_i d \leq \sum_{l \in I_i} r_l - \sum_{k \in O_i} r_k, \quad i \in \Gamma \tag{8c}$$

$$\sum_{B \subset \Lambda_l} x_l^{B,m} \leq y_m, \quad l \in \Lambda, m \in M \tag{8d}$$

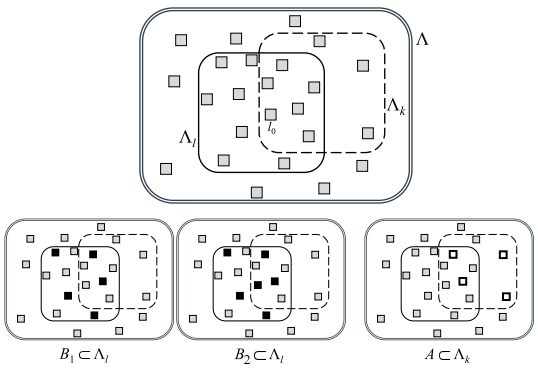


Fig. 3. Examples of consistent and inconsistent local patterns: local pattern B_1 of link l (left) is compatible with local pattern A of link k (right), while local pattern B_2 (middle) is not. Using the binary activation variables, the conflict between B_2 and A is enforced by imposing the constraint $q_l^{B_2,m} + q_k^{A,m} \leq 1$. Gray squares correspond to inactive links, filled black squares correspond to links active in local patterns of link l, and white squares with black outlines correspond links active in local patterns of link k.

$$\sum_{m \in M} y_m \leq 1, \tag{8e}$$

$$x_l^{B,m} \leq q_l^{B,m}, \quad l \in \Lambda, B \subset \Lambda_l, m \in M \tag{8f}$$

$$q_l^{B,m} + \sum_{\substack{A \subset \Lambda_k \\ B \cap \Lambda_k \neq A \cap \Lambda_l}} q_k^{A,m} \leq 1, \tag{8g}$$

$$l, k \in \Lambda, B \subset \Lambda_l, m \in M$$

$$q_l^{B,m} \in \{0,1\}, \quad l \in \Lambda, B \subset \Lambda_l, m \in M \tag{8h}$$

$$x_l^{B,m} \geq 0. \quad l \in \Lambda, B \subset \Lambda_l, m \in M \tag{8i}$$

In this formulation, r_l is the data rate of link l and (8c) is the set of flow constraints that guarantees a minimum downlink throughput of $\alpha_i d$ to non-gateway node i. Equalities have been relaxed to inequalities in the rate, flow, and resource constraints, (8b), (8c), and (8e). Note that many of the consistency constraints of (7) have been bundled into a single inequality in (8g); this is possible because only one local pattern is active for each link in each slot, meaning these constraints can be compounded for local patterns of a single link.

The allocation that emerges from solving this problem is constructed as follows. The m-th global activation pattern is obtained by

$$P_m = \bigcup_{l \in \Lambda, B \subset \Lambda_l : q_l^{B,m} = 1} B,$$

and is alloted a time slot of length y_m normalized to the total frame. Using Theorem 1, it can be shown that any solution to (3) has an N-sparse equivalent that is further equivalent, in terms of global patterns and allocations, to a solution of (8). We omit the proof and refer the reader to [39] for the proof technique.

The objective function in (8) can also be modified to incorporate delay penalization, by rewriting it as

$$d - \lambda \sum_{l \in \Lambda} r_l = d - \lambda \sum_{m \in M} \sum_{l \in \Lambda} \sum_{B \subset \Lambda_l : l \in B} x_l^{B,m} \gamma_l^B.$$

The second term is the sum of data rate on all links, and the weighting factor λ is chosen with the same threshold as derived for the combinatorial formulation in (5).

Both the objective and constraints of this formulation are linear, while the optimization variables are a mixture of continuous and discrete (binary) variables. Thus the exponentially growing linear program of (3) is reduced to a polynomially scaling mixed integer linear program. The number of variables in this formulation grows polynomially with network size due to the fact that as network size grows, *neighborhood* sizes remain the same. If the neighborhood size is no greater than μ, the number of local variables $x_l^{B,m}$ will be no more than $N \times L \times 2^{\mu}$. This brings the number of variables in (8) to a total of fewer than $N \times L \times 2^{\mu} + N + L + 1$ continuous and $N \times L \times 2^{\mu}$ integer values.

Similar to the combinatorial formulation, the above problem can also be modified to include both downlink and uplink by maintaining different link rate variables utilized for the two directions of service. The resulting allocation problem is formulated below.

$$\underset{[x_l^{B,m}],[q_l^{B,m}],[y_m],[r_l^d],[r_l^u],c}{\text{maximize}} \quad c \tag{9a}$$

$$\text{subject to } r_l^d + r_l^u = \sum_{m \in M} \sum_{B \subset \Lambda_l : l \in B} x_l^{B,m} \gamma_l^B, \quad l \in \Lambda \tag{9b}$$

$$\sum_{l \in I_i} r_l^{\mathrm{d}} - \sum_{k \in O_i} r_k^{\mathrm{d}} \geq \alpha_i c, \quad i \in \Gamma \tag{9c}$$

$$\sum_{l \in O_i} r_l^{\mathrm{u}} - \sum_{k \in I_i} r_k^{\mathrm{u}} \geq \beta_i c, \quad i \in \Gamma \tag{9d}$$

$$\sum_{B \subset \Lambda_l} x_l^{B,m} \leq y_m, \quad l \in \Lambda, m \in M \tag{9e}$$

$$\sum_{m \in M} y_m \leq 1, \tag{9f}$$

$$x_l^{B,m} \leq q_l^{B,m}, \quad l \in \Lambda, B \subset \Lambda_l, m \in M \tag{9g}$$

$$q_l^{B,m} + \sum_{\substack{A \subset \Lambda_k \\ B \cap \Lambda_k \neq A \cap \Lambda_l}} q_k^{A,m} \leq 1, \tag{9h}$$

$$l, k \in \Lambda, B \subset \Lambda_l, m \in M$$

$$q_l^{B,m} \in \{0,1\}, \quad l \in \Lambda, B \subset \Lambda_l, m \in M \tag{9i}$$

$$x_l^{B,m} \geq 0, \quad l \in \Lambda, B \subset \Lambda_l, m \in M \tag{9j}$$

$$r_l^d \geq 0, \quad r_l^u \geq 0. \quad l \in \Lambda \tag{9k}$$

Although the localized formulation is scalable in terms of problem size, unlike (4) (which is a linear program), this reformulation requires solving a mixed integer (binary) linear program which is inherently a combinatorial problem and NP-hard. While effective techniques have been devised for solving such problems, there are no guarantees for their computational efficiency. In the next section, some observations regarding the scalability and behavior of the two optimization frameworks of (3) and (8) are discussed.

V. Results and Discussion

Solving the linear program formulated in (3) is relatively straightforward using standard techniques such as the simplex method. Standard solvers such as the CVX package in Matlab or Gurobi were used to solve this problem for small networks of up to 15 links. However, due to exponential growth of the problem size, solving a network with more than twenty links is practically impossible. In [43], this problem is sidestepped by clustering the network around gateways and solving each cluster independently. This provides a suboptimal solution wherein heuristics must be employed to determine cluster association.

The scalable formulation of (8), on the other hand, is a non-convex optimization problem due to the presence of integer variables. Approaches such as the branch-and-bound algorithm are generally effective methods for solving integer linear programs, but provide no guarantees for computational efficiency . For moderately large problems such as the 4-gateway network of Fig. 1(a) , solving (8) using standard solvers (such as the CVX package in Matlab or the Gurobi solver) takes an impractical amount of time. However, we find that in practice, due to sparsity of the interference matrix, the number of global patterns that are activated in the solution can be much smaller than the upper bound guaranteed by Theorem 1, so that the number of global time slots can be set to a value $T \ll N$, indexed by the truncated set $M_T = \{1, \ldots, T\}$. Choosing an appropriate truncation level, T, involves a trade-off between computation time and throughput, as discussed in the next section.

A. Computational Efficiency and Scaling

Limiting the number of global activation patterns to a small number significantly reduces the computation time while achieving virtually all of the optimal throughput. It is also attractive in terms of implementation. Fig. 4 depicts this trend, showing runtime and backhaul throughput obtained from solving (8) with different values for the number of global slots, T, in the urban network of Fig. 1(a) as well as randomly generated suburban networks of 50 and 100 nodes similar to the structure depicted in Fig. 1(b) with gateway density $\eta = 10\%$. Note that the networks used in these simulations are randomly generated. Actual designed networks may have better characteristics in terms of node and gateway placement, as well as better choice of neighbor association, that prevent throughput bottlenecks and produce a more connected network. This would result in more uniform distribution of service among nodes and a higher max-min optimum throughput.

Fig. 5 shows runtime as a function of network size for the original and scalable formulation (with the number of global slots, T, limited to 4 for the latter). Networks of different sizes are generated using the suburban environment model. Formulation (3) blows up exponentially with network size, whereas the BLP of (8) can be used to optimize networks of up to hundreds of nodes within a time-scale of minutes. In running the optimization, we found that by relaxing the rate equation of (8b) to inequality ($r_l \leq \sum_m \sum_B x_l^{B,m} \gamma_l^B$) computation

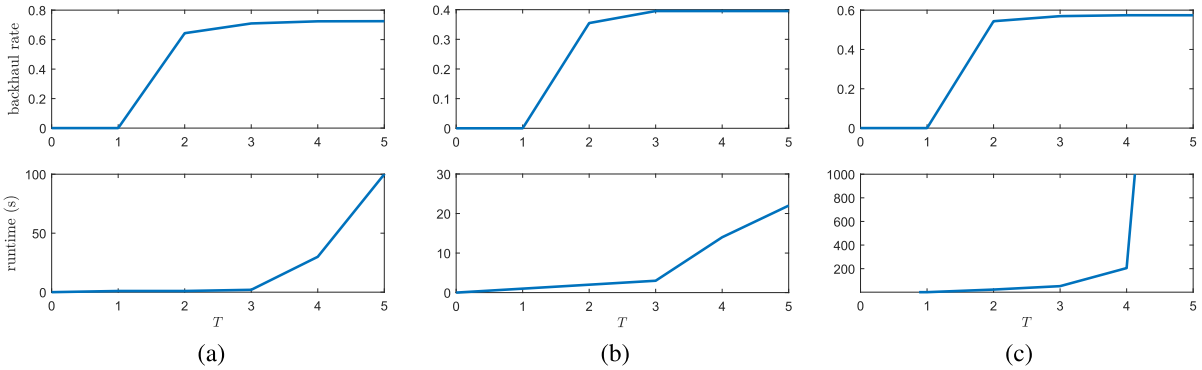


Fig. 4. Effect of truncating M on obtained throughput and computation time in (a) a 4-gateway urban network with 48 nodes, (b) a 4-gateway suburban network with 50 nodes, and (c) a 9-gateway suburban network with 100 nodes. Nominal link rate is 3.46 (SNR=10dB). Depending on network topology, the minimum backhaul data rate delivered to every node is between 10% and 20% of nominal link rate.

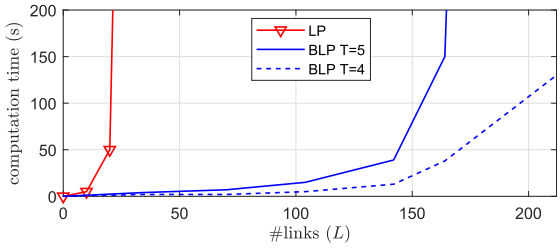


Fig. 5. Computation time as a function of network size for combinatorial and truncated localized formulation.

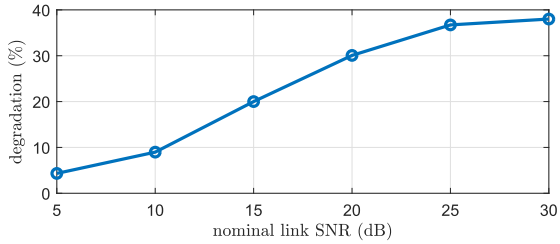


Fig. 6. Degradation of throughput due to suboptimal interference-agnostic scheduling, as a function of nominal link SNR (typical example; numbers derived by evaluating (8) on the urban network of Fig. 1(a)).

time decreased considerably. The results presented here were derived with this relaxation.

B. Effect of Residual Interference

To quantify the effect of interference on backhaul capacity, we compare the throughput obtained in the two cases of (a) only modeling the half-duplex constraint (collocated TX-RX interference), and (b) including the full interference model described in Section II, for the urban network of Fig. 1(a). We find that, similar to results reported in [37] and [43], the optimal throughput capacity is the same in both cases. In fact, the half-duplex nature of transmissions requires that any highly utilized (bottleneck) link at most be activated for a fraction of the time, so that the transmitted data can be *relayed* on the next link(s) on the multihop path in the remainder of the frame. This redundancy in link activation provides room for scheduling links such that no two interfering links are activated simultaneously. Thus backhaul capacity is not degraded from interference, but can only be obtained by careful scheduling of interfering links, and including interference in the model used for solving the allocation problem is crucial for obtaining such a schedule. In fact, if an interference-agnostic allocation is deployed in the network of Fig. 1(a), the resulting max-min backhaul throughput is degraded by around 20% when evaluated in the presence of interference. This degradation becomes more severe as link SNR increases, as depicted in Fig. 6.

C. Backhaul Capacity in Downlink and Uplink

The backhaul throughput provided by the network differs depending on the density of gateway nodes, number of backhaul links connected to each gateway, network structure, and directionality of antennas. For a 10:1 ratio between non-gateway and gateway nodes, approximately 20% of the nominal link data rate can be delivered to nodes as downlink backhaul. Comparing with an existing network in which every base station is directly wired to the Internet, mm-wave wireless backhaul enables 10X shrinking of cell sizes by adding non-gateway base stations, resulting in significantly improved spatial reuse. The immense capacity of densely deployed picocellular access points predicted in [3] can thus be realized using wireless mesh backhauling, as long as the backhaul link data rates are high enough. For 10X increase in access point density, LTE cells with cell traffic in the order of hundreds of Mbps (or 1 Gbps with carrier aggregation) can be supported using wireless links with data rate of several Gbps, which is possible using the unlicensed 60 GHz band. On the other hand, high speed picocells that provide multi-Gbps mmwave-to-mobile access links may require tens of Gbps of backhaul throughput. This, in turn, would require backhaul links with raw data rates of the order of 100 Gbps, which could be realized using mm-wave or THz bands above 100 GHz.

To quantify the joint downlink and uplink throughputs, we fix the downlink ratios α_i to unity and solve for different values of uniformly distributed uplink ratios $\beta_i = \beta$. We observe that an uplink ratio of up to $\beta = 0.6$ can be supported with less than 5% degradation of downlink throughput. The trade-off between downlink and uplink rate is depicted in Fig. 7 for the urban network of Fig. 1(a). Providing equal uplink and downlink capacity results in a throughput reduction of only 20% relative to downlink-only support. This

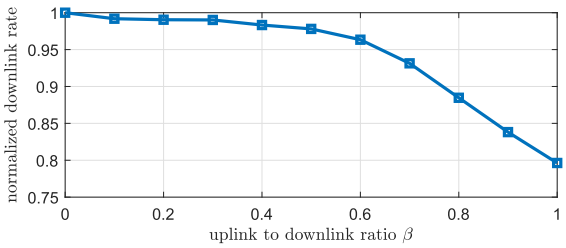


Fig. 7. Reduction in downlink service rate as a function of uplink to downlink ratio. Results based on solving the urban network of Fig. 1(a) with $T = 4$ global patterns.

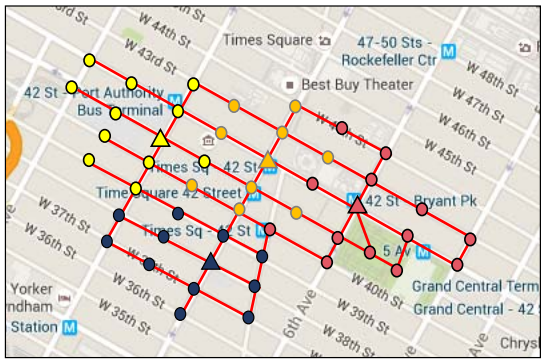


Fig. 8. Example clusters formed by associating links to nearest gateway.

is expected; the link capacity that is idle because of half-duplex relaying is effectively utilized by the uplink traffic that flows in the opposite direction, without much interference on the links carrying downlink traffic.

D. Cost of Clustering

To demonstrate the benefit of optimizing a large multi-gateway network instead of independent optimization of single-gateway clusters, we manually divided the network of Fig. 1(a) into four clusters as shown in Fig. 8. Solving the entire 4-gateway network of Fig. 1(a) results in per-node downlink rate of 21% nominal link rate, whereas solving for the clusters depicted in Fig. 8 independently provides throughput of 17% to each node. Thus as network size grows, optimizing the network directly is preferable to clustering and provides higher throughput.

VI. Conclusion

In this paper, a multihop mm wave mesh network has been proposed for wireless backhaul of urban and suburban picocells. For joint resource scheduling and routing, a scalable formulation in the form of a mixed integer linear program was constructed that is able to solve moderately large networks in a time scale of minutes using a standalone PC; fast enough to adapt to slow varying traffic and topology variations. We observed that using the high speed backhaul links and angular isolation realized by mm wave antenna arrays, an order of magnitude increase in base station density can be supported. We also demonstrated that the formulation can be extended to jointly optimize uplink and downlink and observed that due to the redundancy caused by half duplex traffic relaying, uplink data can be routed without degradation to downlink throughput. Simulation results were reported for urban and suburban environment models, with interference models that account for building blockage and antenna patterns in the respective scenarios.

We note that the proposed infrastructure, while designed here for the task of providing picocellular backhaul, can more generally provide an efficient and secure backbone for a myriad of applications for "smart cities" and metropolitan-wide internet-of-things, such as real-time traffic monitoring and control, support for vehicular communication and autonomy, surveillance, and public transportation, to name a few. Extension of our optimization framework to incorporate a blend of applications with different levels of delay sensitivity is an interesting topic for exploration.

Possible future directions of research also include speeding up the scheduling procedure using iterative optimization approaches or dynamic programming, possibly at the cost of slightly suboptimal solutions. Jointly optimizing the placement of nodes and neighbor associations can also improve backhaul performance.

References

[1] X. Ge, H. Cheng, M. Guizani, and T. Han, "5G wireless backhaul networks: Challenges and research advances," *IEEE Netw.*, vol. 28, no. 6, pp. 6–11, Nov./Dec. 2014.

[2] Y. Zhu *et al.*, "Demystifying 60GHz outdoor picocells," in *Proc. 20th Annu. Int. Conf. Mobile Comput. Netw.*, New York, NY, USA, 2014, pp. 5–16. doi: 10.1145/2639108.2639121.

[3] Z. Marzi, U. Madhow, and H. Zheng, "Interference analysis for mm-Wave picocells," in *Proc. IEEE Global Commun. Conf.*, Dec. 2015, pp. 1–6.

[4] S. Singh, M. N. Kulkarni, A. Ghosh, and J. G. Andrews, "Tractable model for rate in self-backhauled millimeter wave cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, pp. 2196–2211, Oct. 2015.

[5] K. Zheng, L. Zhao, J. Mei, M. Dohler, W. Xiang, and Y. Peng, "10 Gb/s hetsnets with millimeter-wave communications: Access and networking—challenges and protocols," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 222–231, Jan. 2015.

[6] R. Taori and A. Sridharan, "Point-to-multipoint in-band mmwave backhaul for 5G networks," *IEEE Commun. Mag.*, vol. 53, no. 1, pp. 195–201, Jan. 2015.

[7] Y. Yi, G. De Veciana, and S. Shakkottai, "On optimal MAC scheduling with physical interference," in *Proc. 26th IEEE Int. Conf. Comput. Commun.*, May 2007, pp. 294–302.

[8] T. ElBatt and A. Ephremides, "Joint scheduling and power control for wireless ad hoc networks," *IEEE Trans. Wireless Commun.*, vol. 3, no. 1, pp. 74–85, Jan. 2004.

[9] P. Gupta and P. R. Kumar, "The capacity of wireless networks," *IEEE Trans. Inf. Theory*, vol. 46, no. 2, pp. 388–404, Mar. 2000.

[10] D. Yuan, H.-Y. Lin, J. Widmer, and M. Hollick, "Optimal joint routing and scheduling in millimeter-wave cellular networks," in *Proc. IEEE Conf. Comput. Commun.*, Apr. 2018, pp. 1205–1213.

[11] G. Narlikar, G. Wilfong, and L. Zhang, "Designing multihop wireless backhaul networks with delay guarantees," *Wireless Netw.*, vol. 16, no. 1, pp. 237–254, 2010.

[12] X. Lin and S. Rasool, "A distributed joint channel-assignment, scheduling and routing algorithm for multi-channel ad-hoc wireless networks," in *Proc. 26th IEEE Int. Conf. Comput. Commun.*, May 2007, pp. 1118–1126.

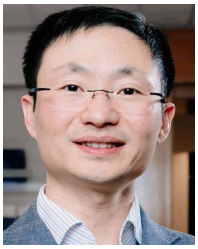
[13] C. Y. Hong and A. C. Pang, "3-approximation algorithm for joint routing and link scheduling in wireless relay networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 2, pp. 856–861, Feb. 2009.

[14] X. Wang and J. Garcia-Luna-Aceves, "Embracing interference in ad hoc networks using joint routing and scheduling with multiple packet reception," *Ad Hoc Netw.*, vol. 7, no. 2, pp. 460–471, 2009.

[15] J. Tang, G. Xue, C. Chandler, and W. Zhang, "Interference-aware routing in multihop wireless networks using directional antennas," in *Proc. IEEE 24th Annu. Joint Conf. IEEE Comput. Commun. Soc.*, vol. 1, Mar. 2005, pp. 751–760.

[16] B. Mumey, J. Tang, and T. Hahn, "Joint stream control and scheduling in multihop wireless networks with MIMO links," in *Proc. IEEE Int. Conf. Commun.*, May 2008, pp. 2921–2925.
[17] J. Zhang, H. Wu, Q. Zhang, and B. Li, "Joint routing and scheduling in multi-radio multi-channel multi-hop wireless networks," in *Proc. 2nd Int. Conf. Broadband Netw.*, Oct. 2005, pp. 631–640.
[18] R. Bhatia and M. Kodialam, "On power efficient communication over multi-hop wireless networks: Joint routing, scheduling and power control," in *Proc. IEEE INFOCOM*, vol. 2, Mar. 2004, pp. 1457–1466.
[19] M. Alicherry, R. Bhatia, and L. E. Li, "Joint channel assignment and routing for throughput optimization in multi-radio wireless mesh networks," in *Proc. 11th Annu. Int. Conf. Mobile Comput. Netw.*, 2005, pp. 58–72.
[20] K. Jain, J. Padhye, V. N. Padmanabhan, and L. Qiu, "Impact of interference on multi-hop wireless network performance," *Wireless Netw.*, vol. 11, no. 4, pp. 471–487, Jul. 2005.
[21] R. L. Cruz and A. V. Santhanam, "Optimal routing, link scheduling and power control in multihop wireless networks," in *Proc. 32nd Annu. Joint Conf. IEEE Comput. Commun.*, vol. 1, Mar. 2003, pp. 702–711.
[22] J. García-Rois *et al.*, "On the analysis of scheduling in dynamic duplex multihop mmwave cellular systems," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6028–6042, Nov. 2015.
[23] B. Li, D. Zhu, and P. Liang, "Small cell in-band wireless backhaul in massive multiple-input multiple-output systems," in *Proc. IEEE Int. Conf. Commun.*, Jun. 2015, pp. 1838–1844.
[24] S. W. Peters and R. W. Heath, Jr, "The future of WiMAX: Multihop relaying with IEEE 802.16j," *IEEE Commun. Mag.*, vol. 47, no. 1, pp. 104–111, Jan. 2009.
[25] Y. Li, J. Luo, W. Xu, N. Vucic, E. Pateromichelakis, and G. Caire, "A joint scheduling and resource allocation scheme for millimeter wave heterogeneous networks," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Mar. 2017, pp. 1–6.
[26] E. Karamad, R. S. Adve, Y. Lostanlen, F. Letourneux, and S. Guivarch, "Optimizing placements of backhaul hubs and orientations of antennas in small cell networks," in *Proc. IEEE Int. Conf. Commun. Workshop*, Jun. 2015, pp. 68–73.
[27] J. Du, E. Onaran, D. Chizhik, S. Venkatesan, and R. A. Valenzuela, "Gbps user rates using mmWave relayed backhaul with high-gain antennas," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 6, pp. 1363–1372, Jun. 2017.
[28] R. Nelson and L. Kleinrock, "Spatial TDMA: A collision-free multihop channel access protocol," *IEEE Trans. Commun.*, vol. 33, no. 9, pp. 934–944, Sep. 1985.
[29] A. Ephremides and T. V. Truong, "Scheduling broadcasts in multihop radio networks," *IEEE Trans. Commun.*, vol. 38, no. 4, pp. 456–460, Apr. 1990.
[30] S. A. Borbash and A. Ephremides, "Wireless link scheduling with power control and SINR constraints," *IEEE Trans. Inf. Theory*, vol. 52, no. 11, pp. 5106–5111, Nov. 2006.
[31] M. Johansson and L. Xiao, "Cross-layer optimization of wireless networks using nonlinear column generation," *IEEE Trans. Wireless Commun.*, vol. 5, no. 2, pp. 435–445, Feb. 2006.
[32] S. Kompella, J. E. Wieselthier, A. Ephremides, H. D. Sherali, and H. D. Sherali, "On optimal SINR-based scheduling in multihop wireless networks," *IEEE/ACM Trans. Netw.*, vol. 18, no. 6, pp. 1713–1724, Dec. 2010.
[33] A. Capone, I. Filippini, and F. Martignon, "Joint routing and scheduling optimization in wireless mesh networks with directional antennas," in *Proc. IEEE Int. Conf. Commun.*, May 2008, pp. 2951–2957.
[34] B. Hajek and G. Sasaki, "Link scheduling in polynomial time," *IEEE Trans. Inf. Theory*, vol. 34, no. 5, pp. 910–917, Sep. 1988.
[35] C.-J. Lin, S.-H. Lin, and C.-F. Chou, "Performance study of optimal routing and channel assignment in wireless mesh networks," in *Proc. IEEE Global Telecommun. Conf.*, Nov. 2007, pp. 4818–4822.
[36] J. Luo, C. Rosenberg, and A. Girard, "Engineering wireless mesh networks: Joint scheduling, routing, power control, and rate adaptation," *IEEE/ACM Trans. Netw.*, vol. 18, no. 5, pp. 1387–1400, Oct. 2010.
[37] M. N. Kulkarni, A. Ghosh, and J. G. Andrews, "Max-min rates in self-backhauled millimeter wave cellular networks," 2018, *arXiv:1805.01040*. [Online]. Available: https://arxiv.org/abs/1805.01040
[38] B. Zhuang, D. Guo, and M. L. Honig, "Traffic-driven spectrum allocation in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, pp. 2027–2038, Oct. 2015. [Online]. Available: http://arxiv.org/abs/1408.6011
[39] Z. Zhou and D. Guo, "1000-cell global spectrum management," in *Proc. 18th ACM Int. Symp. Mobile Ad Hoc Netw. Comput.*, Jul. 2017, p. 20.
[40] Q. Kuang, W. Utschick, and A. Dotzler, "Optimal joint user association and multi-pattern resource allocation in heterogeneous networks," *IEEE Trans. Signal Process.*, vol. 64, no. 13, pp. 3388–3401, Jul. 2016.
[41] B. Zhuang, D. Guo, and M. L. Honig, "Energy-efficient cell activation, user association, and spectrum allocation in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 823–831, Apr. 2016.
[42] B. Zhuang, D. Guo, E. Wei, and M. L. Honig, "Large-scale spectrum allocation for cellular networks via sparse optimization," *IEEE Trans. Signal Process.*, vol. 66, no. 20, pp. 5470–5483, Oct. 2018.
[43] M. E. Rasekh, D. Guo, and U. Madhow, "Interference-aware routing and spectrum allocation for millimeter wave backhaul in urban picocells," in *Proc. 53rd Annu. Allerton Conf. Commun., Control, Comput.*, Sep./Oct. 2015, pp. 1–7.
[44] G. Meurant, *Handbook of Convex Geometry*. Amsterdam, The Netherlands: Elsevier, 2014.
[45] M. Pióro and D. Medhi, *Routing, Flow, and Capacity Design in Communication and Computer Networks*. Amsterdam, The Netherlands: Elsevier, 2004.

Maryam Eslami Rasekh received the B.S. degree in electrical engineering from the Isfahan University of Technology in 2007 and the M.S. degree from the Sharif University of Technology in 2009. She is currently pursuing the Ph.D. degree in electrical and computer engineering with the University of California Santa Barbara. Her research interests include next-generation wireless communication and sensing, and signal processing for large scale MIMO systems.



Dongning Guo (S'97-M'05-SM'11) received the Ph.D. degree from Princeton University, Princeton, NJ. He then joined the faculty of Northwestern University, Evanston, IL, where he is currently a Professor in the Department of Electrical and Computer Engineering. He has been an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION THEORY and a Guest Editor of a Special Issue of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS. He is an Editor of *Foundations* and *Trends in Communications and Information Theory* and an Associate Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS. He was a recipient of the IEEE Marconi Prize Paper Award in Wireless Communications in 2010 and a Best Paper Award at the 2017 IEEE Wireless Communications and Networking Conference. He is also the recipient of the National Science Foundation Faculty Early Career Development (CAREER) Award in 2007.

Upamanyu Madhow is a Distinguished Professor of Electrical and Computer Engineering at the University of California, Santa Barbara. His current research interests focus on next generation communication, sensing and inference infrastructures centered around millimeter wave systems, and on robust machine learning. He received his bachelor's degree from the Indian Institute of Technology, Kanpur, in 1985, and his Ph. D. degree from the University of Illinois, Urbana-Champaign, in 1990, all in electrical engineering. He has worked as a research scientist at Bell Communications Research, Morristown, NJ, and as a faculty at the University of Illinois, Urbana-Champaign. He is a recipient of the 1996 NSF CAREER award, and co-recipient of the 2012 IEEE Marconi prize paper award in wireless communications. He has served as an Associate Editor for the IEEE TRANSACTIONS ON COMMUNICATIONS, the IEEE TRANSACTIONS ON INFORMATION THEORY, and the IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY. He is the author of two textbooks published by Cambridge University Press, *Fundamentals of Digital Communication* (2008) and *Introduction to Communication Systems* (2014).