On the Analysis of Scheduling in Dynamic Duplex Multihop mmWave Cellular Systems

J. García-Rois*, F. Gómez-Cuba*, M.R. Akdeniz†, F. J. González-Castaño*, J. C. Burguillo-Rial*, S. Rangan†, B. Lorenzo*,

Abstract

With the shortage of spectrum in conventional cellular frequencies, millimeter-wave (mmWave) bands are being widely considered for use in next-generation networks. Multihop relaying is likely to play a significant role in mmWave cellular systems for self backhauling, range extension and improved robustness from path diversity. However, designing scheduling policies for these systems is challenging due to the need to account for both adaptive directional transmissions and dynamic time-division duplexing schedules, which are key enabling features of mmWave systems. This paper considers the problem of joint scheduling and congestion control in a multihop mmWave network using a Network Utility Maximization (NUM) framework. Interference is modeled with an exact model and two auxiliar simplified models: actual interference (AI), with a graph-based calculation of the Signal to Interference plus Noise Ratio (SINR) depending on dynamic link activity and directivity, as well as upper and lower bounds computed from worst-case interference (WI) and interference free (IF) approximations. Throughput and utility optimal policies are derived for all interference models (AI, WI and IF) with both deterministic Maximum Weighted and randomized Pick and Compare scheduling algorithms, jointly with decentralized Dual Congestion Control. Results are evaluated with numerical simulations, using accurate mmWave channel and beamforming gain approximations based on measurement campaigns.

Index Terms

5G, millimeter-wave, beamforming, dynamic duplexing, scheduling, utility maximization

*AtlantTIC, University of Vigo, 36310 Vigo, Spain. {jgrois, fgomez, javier, jrial, blorenzo}@gti.uvigo.es.†NYU Polytechnic Scool of Engineering, Brooklyn, NY, USA. {akdeniz, srangan}@nyu.edu. This research has been funded by project Táctica, MINECO, Spain, the National Science Foundation under Grants 1116589 and 1237821 and generous support from NYU WIRELESS affiliates.

I. Introduction

A. Motivation

Demand for cellular wireless data has been growing rapidly [1]. With the severe shortage of spectrum in conventional cellular frequencies below 3 GHz, millimeter-wave (mmWave) bands, roughly between 30 and 300 GHz, are now being widely considered for next-generation systems [2]–[5]. These bands have significant differences with respect to current micro-wave (μWave) systems beyond the large amount of new spectrum. On the one hand, the higher atmospheric absorption of mmWave greatly reduces received power and has been a major drawback so far. On the other hand, shorter wavelengths allow large antenna arrays to fit into current hand-held devices. These arrays implement highly directional beamforming gains that alleviate the higher pathloss of mmWave. While propagation outside the main transmission directions cause interference below the noise level, the interference generated by the main transmission directions may be avoided by appropriate scheduling on a case-by-case basis. Fully directive communications with little interference are the key enabler of highly flexible resource allocation mechanisms in mmWave involving time, frequency and space domains. Conversely, in current μWave cellular systems, there is simultaneous interference from many directions due to omnidirectional transmissions, and hence more static subframe structures are required.

Even though directive transmissions may overcome the higher pathloss of mmWave bands, solid obstacles may cause outages. In fact, recent measurement campaigns in [6] have led to a three-state channel model with LOS, NLOS and outage probabilities that we apply in our work. Mass deployment of access points (APs) is necessary to prevent these outages, and this requires a wireless multihop (*relaying*) architecture in scenarios where fiber is not available or is too expensive. Such architecture also provides range extension and path diversity against intermittent outages in the mmWave range. A recent theoretical analysis in [7] also suggests that, in very wideband regimes (as in the case of mmWave), multihop communications are necessary to fully exploit all available degrees of freedom. Nonetheless, a key challenge in realizing the full benefits of multihop relaying is the appropriate design of dynamic scheduling policies and Congestion Control (CC) mechanisms.

μWave relaying in the context of 3GPP Long Term Evolution (LTE) [8], [9] has achieved limited benefits in throughput improvement and is mainly considered for coverage extension. The main reason for this is the global *static duplexing* scheme in current time-division duplex

(TDD) cellular systems. This scheme requires all cells to perform either downlink (DL) or uplink (UL) transmissions and is particularly troublesome for relay nodes (RNs), as they cannot appropriately balance two-hop link resources and experience significant bottlenecks [10]. Static duplexing is necessary in traditional μWave systems due to their inherent interference-limited nature, i.e. they are Signal-to-Interference-plus-Noise Ratio (SINR) based networks. However, due to their high directivity, mmWave networks have been observed to be noise-limited [6], [11]–[13]. This allows full dynamic resource allocation in TDD mmWave systems – a technique known as *dynamic TDD* [14], where the duplexing schedule at each frame can be adaptively adjusted on a per link basis with no constraints on UL/DL operation.

This work studies joint scheduling and CC for a general class of mmWave networks with a multihop backhauling mesh structure. The network consists of base stations (BSs), relay nodes (RNs) and users (UEs), with both UL and DL traffic flows along arbitrary routes. Each node is subject to two fundamental scheduling constraints: *half-duplex* communication, so that nodes cannot transmit and receive at the same time, and *one-to-one* communication, so that a node cannot use several links at a time, since analog beamforming is expected to be limited to one direction at a time in the first generation of mmWave devices [15]. Dynamic duplexing consists on link scheduling decisions, which are made on a slotted basis. The Actual Interference (AI) is computed using a graph-based SINR model. Because spatial isolation reduces the importance of interference and power control in mmWave, we opted for a simplified power allocation strategy, and show in our results that optimal power control is unlikely to provide a significant throughput improvement.

Our main contribution is a joint scheduling and CC policy that achieves maximum capacity under fairness requirements within the AI model. In order to maximize the set of feasible UL/DL rates per user and enforce fairness, we follow the classic Network Utility Maximization (NUM) framework. Given that the AI model involves coupling across several layers of the protocol stack, including the Physical (PHY), Medium Access Control (MAC), network, and transport layers, the main problem is addressed using a cross-layer decomposition. This allows us to approach the classic optimal solution of Maximum Weighed Scheduling (MWS), based on *differential backlogs* over *variable capacity* links, using also the classic Pick-and-Compare (PaC) randomized scheduling algorithm with a Dual CC (DCC) mechanism. In order to show that throughput-utility optimality is achieved for the AI model, we adapt the proof in [16] to the case in which link capacity experiences accurately calculated schedule-dependent mmWave interference.

Unfortunately, the variation of queues requires MWS to be computed each frame, even with constant link capacities across the time, which is a very time consuming task; this problem becomes NP-hard each frame when the interference is scheduling-dependent, as it is in the AI model. Considering recent work in mmWave that assumed negligible interference [11]–[13], we also define simplified (but still *topology-aware*) Interference Free (IF) and Worst-case Interference (WI) models. These models determine upper and lower bounds for the capacity achieved by the AI model and are computationally simpler due to the assumption of static link capacities. Our simulations under the AI model indicate that there are scheduling instances where interference by simultaneously selected links significantly affects the capacity of a number of those links. However, with our throughput-utility optimal network management, the simplified IF model remains a very tight upper bound for the capacity in the AI results. This occurs because optimal management prevents the scheduling of links with significant cross-talk, as a form of implicit interference avoidance, suggesting that the simplified IF model allows for a realistic evaluation of system capacity in mmWave cellular networks as long as we can guarantee that the real network is operated with optimal throughput control. This represents a fundamental difference with respect to traditional cellular and wireless systems, where a similar IF model would be very inaccurate even with optimal control. A byproduct of this tightness is that a significant part of ad-hoc and wireless graph-theory approaches may be applied realistically to the new mmWave cellular paradigm. Traditional research areas pertaining to μWave cellular networks, such as interference cancellation or power control, are less relevant in the mmWave context.

The algorithms are validated by numerical simulations, using mmWave channel and beamforming models in [6] derived from the *real-world* New York City measurements in [3], [17]. Similar measurements have also been reported in [18].

The paper is organized as follows: In Section II we describe our mmWave network architecture model. In Section III, we formulate the NUM problem and provide a solution for the PaC algorithm with AI, which is the more general case and implies convergence also with the IF and WI models. In Section IV we present numerical simulation results illustrating the optimality of PaC under the AI model (IV-A), in which we compare the three models (AI, IF and WI) in IV-B and the capacity of mmWave cells without RNs and with our multihop architecture (IV-C). Finally, Section V summarizes the contributions of our model and describes future research lines.

B. Related Work

Previous analyses of self-backhaul, multihop relay mmWave networks have focused on either stochastic geometric analyses [19], [20] or scaling laws [7]. These analyses characterize the overall capacity as a function of the key system parameters such as base station and relay node density, but do not explicitly optimize the scheduling policy. There has also been significant work in multihop protocols for mmWave local area systems [21], [22]. Two other aspects of mmWave communication that are critical to the design of multihop systems but are not considered here are cell discovery [23] and channel estimation [24], [25].

The theoretical framework for this paper is derived from Kelly's work in [26], [27], which introduced the NUM approach. In the seminal work [28], the concept of imperfect scheduling was first introduced, permitting sub-optimal instantaneous schedules at each decision point without violating long-term throughput optimality (i.e. PaC). Subsequent extensions over the years introduced multihop [16], [29]; reduced time complexity of control signaling with or without trade-offs [30]; analyzed other QoS metrics beyond throughput and fairness [31], [32]; considered fairness in heterogeneous networks with time-varying channels [33]; dealt with optimal power allocation [34], [35]; and studied the effect of reconfiguration delays on network capacity [36]. We refer readers to [37], [38] for surveys on these topics. With the exception of some recent studies that cover limited models, none of the work to date has considered beamforming or mmWave communications. For instance, the recent work on video quality in [39] does not take interference into account, provides a centralized solution, and only considers indoor scenarios. We provide a random PaC scheduling algorithm and a distributed CC mechanism, which are indirectly aware of the interference between links, and still reach the maximum throughput-utility that any policy could obtain.

Many of the works mentioned above employ the family of K-hop interference models, in which links can be active only if they lie within a K-hop distance to other active links in the topology graph that represents the network. Thus, interference is considered merely as a rule that prevents some simultaneous transmissions. Furthermore, the available capacity in a link is always considered to be the same, regardless of the pairs of transmitters and receivers chosen. This approach thus ignores inter-link interference between simultaneous transmissions (i.e., SINR), but it is very common in the literature because the scheduling problem alone is NP-hard in general.

Whether or not K-hop interference models deliver good approximations depends on the type of network. For instance, a 1-hop interference model is excessively optimistic for general omnidirectional wireless networks, for which higher degrees (2-hop, 3-hop, etc.) are more reasonable. Unfortunately, as K increases, the results underestimate achievable capacity, given the spatial multiplexing loss due to the fewer concurrent links allowed.

However, the 1-hop schema may provide a reasonable approximation for calculating the maximum capacity of mmWave systems, given their higher spatial isolation. We therefore use a 1-hop schema to model half-duplex and one-to-one communications, avoiding what we call throughout this paper *hard interference*, i.e., interference conditions that prevent two links from being simultaneously active, which is a typical consideration in the scheduling literature. Nevertheless, on top of hard interference, we formulate the capacity of the scheduled links as a function that depends on actual SINR, which affects link rates without blocking them completely. Hereafter we will use the term *soft interference* to refer to the change in SINR of a link caused by other links that may be simultaneously active, as is usually considered in physical layer research. Our AI model is, therefore, a new mmWave *hybrid* interference model that combines a 1-hop scheduling constraint and computation of SINR for simultaneously active links.

Even though in [16] it was demonstrated that PaC converges to a set of average rates that are arbitrarily close to the optimum fair allocation under general graph-based interference models, the authors only considered the hard interference constraints of a mmWave network, assuming fixed link capacities. Furthermore, even though Lee et al demonstrated PaC throughput-optimality for SINR interference models in [34], they considered omnidirectional transmissions, which do not allow an accurate representation of mmWave directional transmissions or soft interference in an mmWave network. In addition, although the result accounted for scheduling stability, it did not reach the full NUM solution by completing it with CC.

As previously mentioned, our simplifications to obtain lower and upper bounds (the WI and IF models) consider constant link capacities. Therefore, they remove the hybrid part of AI, leaving only a pure 1-hop hard interference constraint, as in traditional scheduling analysis.

A related consequence of highly directional transmissions is a higher degree of spatial link isolation. In other words, interference has a much weaker effect than in current small-cell networks. Many recent works even assume that inter-link interference levels will be negligible in mmWave networks, and that links can be approximated as *pseudo-wires* [11]–[13], [39]. In any case, some interference mitigation technologies of the last decade, such as coordinated multi-

point, inter-cell interference coordination and interference alignment, will lose the relevance they had in traditional cellular systems [5].

We think that the question of whether the fully isolated pseudo-wired hypothesis is realistic or not in certain scenarios remains open. If transmissions were truly spatially isolated, our three interference models would converge to a plain IF model, and significant results from graph-oriented scheduling research would be directly applicable to the realistic modeling of mmWave networks. However, if soft interference happens to be non-negligible, the accuracy of 1-hop models would be low, and our hybrid model would be essential. This paper does not assume the pseudo-wired hypothesis. We employ the mmWave AI model with accurate schedule-dependent representation of interference to derive our main result. Nonetheless, we do compare the WI, AI and IF models to discuss the accuracy of the hypothesis.

II. System model

A. Network, connections, and traffic flows.

We built our model on the assumption that we avail of a sensing stage [23] in which each node n (a BS, RN or UE) detects and identifies its set of neighbors $\Omega(n)$. We also consider that all nodes remain at fixed locations for the period of interest of the scheduling algorithm. We represent the network as a directed graph $\mathcal{G}(\mathcal{N}, \mathcal{L})$, where $\mathcal{N}$ is the set of nodes (BSs, RNs and UEs), $\mathcal{L}$ is the set of links, indexed by n and ℓ respectively, and $\mathcal{F}$ is the set of traffic flows in the network, indexed by f. We denote the cardinalities of these sets as N, L and F. Admissible connections are UE⇌RN, UE⇌BS, RN⇌RN and RN⇌BS. Furthermore, due to the half-duplex and one-to-one restrictions, only links without shared devices (i.e. edges with no shared vertexes on the graph) can be active at the same time.

Dynamic duplex is modeled as synchronized discrete time frames (indexed by t), in which transmissions between any valid pair of nodes can be allocated. For each pair of nodes that form a link $\ell = (n, m)$, $n, m \in \mathcal{N}$, we define the logic indicator $\pi_{n\to m}(t) = 1$ if node n transmits towards node m at frame t, and $\pi_{n\to m}(t) = 0$ otherwise. Hence, the set of link pairs allocated at frame t is represented by the binary vector $\boldsymbol{\pi}(t)$, also called *a schedule*[1]. The half-duplex and one-to-one communication constraints limit the set of feasible schedules, denoted by Π

[1]Note that in this terminology schedule $\boldsymbol{\pi}(t)$ is simply the set of active links for frame t, and a *scheduling policy* is the complete method for choosing schedules $\boldsymbol{\pi}(t)$ across all time frames in order to solve the desired *scheduling problem*.

with cardinality $\mathcal{M}$. In our model, Π represents the *hard-interference* constraint according to the 1-hop interference rule.

We consider that UL flows are absorbed by BSs, and that each user can also receive data flows from any BS (DL). We also define the set $\mathcal{D} \subset \mathcal{N}$ of feasible destinations, comprising all BSs and UEs, which coincides with the set of feasible sources $\mathcal{S}$, as both BSs and UEs can act as sources or destinations. Flow sources generate packets according to one of the following distributions, depending on whether CC is present:

Definition 1. An *inelastic* packet arrival process by flow f in source node s is an i.i.d. stochastic process with a constant mean arrival rate λ_s^f bits/slot, where in each frame a random number of packets $a_s^f(t)$ is generated, satisfying $E[a_s^f(t)] = \lambda_s^f$, where the process has finite first and second moments.

Definition 2. An *elastic* packet arrival process by flow f in source node s is a stochastic process with a controllable mean arrival rate injected into the network $a_s^f(t)$, with a long-term mean arrival rate $x_s^f = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} a_s^f(t)$.

Note that we commit a minor notation abuse by using $a_s^f(t)$ for both inelastic and elastic arrival rates within a frame. The notion of elastic traffic permits the definition of a control system to both avoid resource under-utilization when the state of the network allows data rate increases and prevent instability when the state of the network requires sources to back-off momentarily. The implementation of the CC combines these two practices.

Finally, we assume *on/off power allocation* for all nodes (BSs, RNs and UEs) for two main reasons. On the one hand, previous research on non-mmWave ad-hoc networks with PaC can be extended to arbitrary power allocation as in [34], but this obscures the analysis without providing additional information, as the technique for integrating power allocation in the throughput-optimality proof consists of modeling random power allocation. On the other hand, several recent works [11]–[13] have assumed a pseudo-wired behavior for mmWave links, and the application of a power control mechanism in such a context is less relevant. Unlike these recent works, we do not assume that interference is negligible, but rather compute real mutual interference between concurrent links within a frame. This allows us to check the accuracy of the pseudo-wired assumption and evaluate the cost of the complexity of power allocation for future works.

B. Directivity, interference and link capacity

A fundamental aspect of our models is that all nodes have adaptive beamforming capabilities as in [18]. This is a realistic assumption given the decreasing costs of circuitry for mmWave frequencies, which will allow high dimensional antenna arrays to fit in a small form factor. This means that the antenna gain for a signal transmitted by a device will depend on the receiver and the intended destination. We formulate the signal received at the destination according to the following discrete channel model:

$$y_m(t) = \underbrace{\mathbf{w}^{\mathrm{r}}_{n\to m}\mathbf{H}_{n\to m}\mathbf{w}^{\mathrm{t}}_{n\to m}g_{n\to m}x_n(t)}_{\text{intended signal}} + \sum_{\substack{d(i),i\neq n,m \\ d(i)\not=\pi_{i\to d(i)}=1}} \underbrace{\mathbf{w}^{\mathrm{r}}_{n\to m}\mathbf{H}_{i\to m}\mathbf{w}^{\mathrm{t}}_{i\to d(i)}g_{i\to m}x_i(t)}_{\text{soft interference}} + z(t) \quad (1)$$

where we denote the macroscopic pathloss and the normalized channel matrix from node n to node m by $g_{n\to m}$ and $\mathbf{H}_{n\to m}$, respectively. Vector $\mathbf{w}^{\mathrm{r}}_{n\to m}$ denotes the receiver beamforming at m when it is expecting a transmission from n; $\mathbf{w}^{\mathrm{t}}_{n\to m}$ denotes the transmitter beamforming at n when it is transmitting towards m; and $i, d(i)$ denotes the interference transmitters and their corresponding destinations. Note that the second term is the *soft interference*, where beamforming vectors are not correctly matched to the channel matrix.

We only consider *single-stream* processing as a simplifying assumption, as explained above, and hence y and x are scalar. We also assume that channel matrices remain constant for the duration of a scheduling interval, and that transmitters design the beamforming vectors to maximize SNR regardless of interference.

$$\mathbf{w}^{\mathrm{r}}_{n\to m}, \mathbf{w}^{\mathrm{t}}_{n\to m} = \arg\max |\mathbf{w}^{\mathrm{r}}\mathbf{H}_{n\to m}\mathbf{w}^{\mathrm{t}}|^2 \quad (2)$$

Since the channel is essentially static and beamforming does not depend on interference in scheduling, a node n can obtain all the necessary beamforming vectors for all the nodes in its neighbor set during the sensing phase ($\mathbf{w}^{\mathrm{r}}_{m\to n}, \mathbf{w}^{\mathrm{t}}_{n\to m}, m \in \Omega(n)$).

For an accurate mmWave link capacity generation as a function of the distance between nodes, we compute the macroscopic pathloss following the distribution in [6]:

$$g_{n\to m} = 75.85 + 37.3\log_{10}(d(n,m)) + \xi, \xi \sim \log\mathcal{N}(0, 8.36) \quad (3)$$

We compute the beamforming gain in the direction of the desired link $n \to m$, denoted as

$$G(n \to m) = |\mathbf{w}^{\mathrm{r}}_{n\to m}\mathbf{H}_{n\to m}\mathbf{w}^{\mathrm{t}}_{n\to m}|^2,$$

$$SINR_{n\to m} = \frac{P_{\max}\pi_{n\to m}(t)G(n\to m)g_{n\to m}}{\sum\limits_{\forall i\neq n}\sum\limits_{\forall j\neq m} G(i\to m|i\to j, n\to m)g_{i\to m}P_{\max}\pi_{i\to j}(t) + WN_0} \tag{4}$$

and the secondary beamforming gain experienced by the interference path $i \to m$ when the active links are $n \to m$ and $i \to d(i)$, which we denote as

$$G(i\to m|n\to m, i\to d(i)) = |\mathbf{w}^{\mathrm{r}}_{n\to m}\mathbf{H}_{i\to m}\mathbf{w}^{\mathrm{t}}_{i\to d(i)}|^2,$$

With these data, the actual capacity of link $n \to m$ at frame t is modeled as

$$c_{n\to m}(t) = \min[\alpha_1 W \log\left(1+\alpha_2 SINR_{n\to m}\right), C_{\max}] \tag{5}$$

where the left-hand side of the minimum represents capacity as a function of SINR and the right-hand side represents the fact that the physical layer has a finite maximum rate $C_{\max} < \infty$. This finite maximum property is usually required in scheduling convergence proofs [34].

In (5), W is always the full mmWave system bandwidth in Hz, because only one link per transmitter is active per node. The power and bandwidth penalty factors $\alpha_1, \alpha_2 < 1$ of the physical layer regarding the Shannon capacity must be obtained from empirical data. For illustration purposes, these values take a $-3dB$ SINR penalty in our simulations. Finally, $SINR_{n\to m}$ is the instantaneous SINR value of link $n \to m$, which is computed based on the active links in the current network schedule $\boldsymbol{\pi}(t)$. This ratio is defined in expression (37), where N_0 is the noise power spectral density and the pathloss and beamforming gains are defined above.

The main difference between mmWave models and the models described in previous literature is high spatial isolation, which causes the average sum of interfering power in the denominator of (37) to shrink, but also results in stronger variations due to changes in scheduling. To evaluate the importance of adapting scheduling to these time-varying capacities in the AI model, we formulated upper and lower network rate bounds using the fixed-capacity IF and WI models.

The IF model is simply the result of removing all soft interference in (37). If the pseudo-wired hypothesis in the literature is correct, this model will be accurate.

$$c^{IF}_{n\to m}(t) = c_{n\to m}(t), \text{ s.t. } G(i\to m|n\to m, i\to d(i)) = 0 \forall (i, d(i)) \neq (n, m) \tag{6}$$

The WI model assumes that all devices transmit at the same time and applies a worst-case beamforming vector for each interfering source,

$$G^{\mathrm{WI}}(i \to m | n \to m) = \max_{\substack{i, d(i) \neq m, n \\ d(i) \in \Omega(i)}} G^{\mathrm{WI}}(i \to m | n \to m, i \to d(i)) \tag{7}$$

which we find through an exhaustive search within its neighbor set $\Omega(i)$. Simultaneous transmissions by all nodes are prohibited by scheduling constraints Π, but they can be represented in a binary vector, defined as $\boldsymbol{\pi}^{\mathrm{WI}n \to m} \notin \Pi$, where active links correspond to the worst beamforming found $\pi_{i \to d(i)}^{\mathrm{WI}n \to m} = 1 \Leftrightarrow i \to d(i) = \arg\max$ (7). Applying this vector to (37), we compute the worst-case link capacities.

$$c_{n \to m}^{WI}(t) = c_{n \to m}(t), \text{ s.t. } \boldsymbol{\pi}(t) = \boldsymbol{\pi}^{\mathrm{WI}n \to m} \tag{8}$$

Note that this worst-case representation differs when calculating the capacity of each link $n \to m$.

As spatial isolation increases ($G(i \to m | n \to m, i \to d(i)) \to 0$), interference tends to zero and the capacities of all three models converge to the IF case, since the pseudo-wired assumption [11] holds. The WI model determined a lower bound on system capacity because AI is strictly always smaller (the worst-case total-interference vector is not contained in the feasible set of vectors allowed by the scheduling rules in Π).

Note that in the AI model we assumed on/off power allocation as a simplification, but for the upper (IF) and lower (WI) bounds, where capacities do not change, the assumption is optimal. Therefore, if both bounds are tight, on/off power allocation is reasonable for the AI model.

III. Network Utility Maximization (NUM)

A. Problem Statement

We consider a queued traffic network where at each node n there are F separate queues for forwarding the packets for each flow. The evolution of the whole set of queues $\mathbf{q}(t) \in \mathbb{N}_0^{N \times F}$ depends on the scheduling policy chosen, as this governs the set of transmitter-receiver pairs allocated at each frame.

Specifically, the evolution of a given queue is given by:

$$q_n^f(t+1) = q_n^f(t) + \left[\sum_{m \in \Omega(n)} \breve{c}_{n \leftarrow m}^f(t) + a_n^f(t) \right] - \left[\sum_{m \in \Omega(n)} \breve{c}_{n \to m}^f(t) \right] \tag{9}$$

where n sustains incoming or outgoing communications with one of its neighbors $m \in \Omega(n)$ and $\breve{c}^f_{n\to m}(t) = \min(c^f_{n\to m}(t), q^f_n)$ is the delivered data rate at time t, constrained by both the remaining packets in the transmission queue and the fact that the dedicated link capacity per flow, $\breve{c}^f_{n\to m}(t)$, must satisfy the total link capacity budget $\sum_f \breve{c}^f_{n\to m}(t) = \breve{c}_{n\to m}(t) \leq c_{n\to m}(t)$. In addition, destination nodes satisfy $q^f_{d(f)} = 0$, where $d(f) \in \mathcal{D}$ is the destination node for flow f, so that data reaching its destination is removed from the network.

We define the stability of the system essentially by requiring that any component $q^f_n(t)$ of the vector $\mathbf{q}(t)$ does not grow to infinity. Hence:

Definition 3. The network is *stable* if the queue lengths at the nodes are bounded [28], i.e. $\lim\limits_{T\to\infty} \sup \frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left[q^f_n(t)\right] < \infty$ for all queues.

The capacity (or stability) region is derived from the previous condition of stability

Definition 4. The *capacity* (or *stability*) *region* Λ is the set of all mean flow rate vectors $\boldsymbol{\lambda}$, or long-term rate vectors $\mathbf{x}$ (for elastic traffic), for which there exists at least one scheduling policy that makes the network stable (i.e., there is a policy that prevents the queues from growing to an infinite size).

Next, we define a throughput-optimal scheduling policy:

Definition 5. A scheduling policy is *throughput-optimal* if it achieves stability for all $\boldsymbol{\lambda} \in \Lambda$ (or $\mathbf{x} \in \Lambda$ for elastic traffic).

Finally, we can define our problem under the NUM framework as follows:

$$\max_{0\leq x^f_n \leq A_{max}} \quad \sum_{n,f} U_{n,f}(x^f_n) \qquad \text{s.t.,} \;\; \mathbf{x} \in \Lambda \tag{10}$$

where $U_{n,f}(x^f_n)$ is the function that measures the utility of achieving an average rate x^f_n on flow f.

The well-known utility framework measures the incentive of allocating a certain rate to each terminal [26], [27]. These objective functions are typically required to be strictly concave, non-decreasing, and twice differentiable [40]. Depending on the utility function chosen, different types of fairness can be achieved (proportional, max-min, etc.). For example, $\log x$ can be used for proportional fairness, as in our simulations.

B. Cross-Layer Decomposition for NUM

We solve the above problem under the AI model to obtain the throughput capacity of the mmWave network. We define the capacity vector $\mathbf{c}(t) = \Phi(\boldsymbol{\pi}(t))$ as the application

$$\boldsymbol{\pi} \in \{0,1\}^{L\times 1} \xrightarrow{\Phi()} \mathbf{c} \in [0, C_{\max}]^{L\times 1} \tag{11}$$

of the PHY layer described in Section II-B, which returns the actual link rates given the schedule vector $\boldsymbol{\pi}(t)$. We define the set of all achievable instantaneous link rate vectors as $\mathbf{c}(t) \in \mathcal{C} = \{\Phi(\Pi)\}$. The constraint $\mathbf{x} \in \Lambda$ in (10) can be expanded as in the following set of rules:

Flow constraint at each node ($\forall f, \forall n \neq d(f)$):

$$\sum_{m\in\Omega(n)} c^f_{n\to m} \geq \sum_{m\in\Omega(n)} c^f_{n\leftarrow m} + x^f_n \tag{12}$$

Interference constraint:

$$\left[\sum_f c^f_{n\to m}\right] \in \text{Co}(\mathcal{C}), \qquad \forall (n\to m) \in \mathcal{L} \tag{13}$$

Non-negative rates:

$$c^f_{n\to m} \geq 0, \qquad \forall (n\to m) \in \mathcal{L} \tag{14}$$

The convex hull $\text{Co}(\mathcal{C})$ of all possible link-capacity values $\mathbf{c}(t) \in \mathcal{C}$ enforces an average rate constraint over the time-average effect of all individual schedules as opposed to just a single schedule. This convex hull guarantees that the region Λ is a convex set [37]. Therefore, the primal problem (10) - (14) has a unique optimum $\overset{*}{\mathbf{x}}$ (i.e. fair allocation) due to the strict concavity of $U_{n,f}(x)$ and the convexity of domain Λ [40].

Based on the duality theory, $\overset{*}{\mathbf{x}}$ can be calculated by solving rate control, routing and scheduling sub-problems separately. In the literature, cross-layer mechanisms have been extensively studied for this purpose [37]. It was shown in [37] that the optimal solution can be found whenever the dual problem of (10) - (14) exists. This requires Λ to be a convex set, as is the case. A Lagrange multiplier is then associated with each rule (12)-(14), and the optimal solution of (10) can be found by solving the following procedures [37], [38]:

Congestion Control [distributed]. The sources find a feasible average traffic data rate $\mathbf{x} \in \Lambda$ with a CC policy on injected backlogged traffic that prevents source queue lengths from growing to infinity. We consider the DCC formula in [40] given by:

$$a^f_n(t) = \min\left\{\max\left[U'^{-1}_{n,f}\left(\frac{q^f_n(t)}{V}\right), 0\right], A_{max}\right\} \tag{15}$$

where $V > 0$ is a constant parameter.

Maximum Weighted Scheduling and Capacity Allocation [centralized]. At each frame t, the network controller chooses $\boldsymbol{\pi}(t) \in \Pi$, such that the corresponding $\mathbf{c}(t)$ under the current queue state $\mathbf{q}(t)$ maximizes the following expression:

$$\overset{*}{\boldsymbol{\pi}}(t) = \underset{\mathbf{c}\in\Phi(\boldsymbol{\pi}),\boldsymbol{\pi}\in\Pi}{\arg\max} \; (\mathbf{c}(t)^T \overset{*}{\mathbf{w}}(t)) \tag{16}$$

where super-index $(\cdot)^T$ denotes vector transposing, and $\overset{*}{\mathbf{w}}(t) \in \mathbb{N}_0^{L\times 1}$ and each component is defined as $\overset{*}{w}_{n\to m}(t) = \max_f(\max(q_n^f(t) - q_m^f(t), 0))$. The components of $\mathbf{c}(t)$ are defined as follows,

$$c_{n\to m}^f(t) = \begin{cases} \Phi_{n\to m}(\boldsymbol{\pi}(t)), & \text{if } \; f = f^* \; \text{ and } \; w_{n\to m}^{f^*}(t) \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

where f^* is the flow that maximizes the differential backlog $w_{n\to m}^{f^*}(t)$ at link $n \to m$.

Queue Update [distributed]. Let $c_{in(n)}^f(t) = \sum_{m\in\Omega(n)} c_{n\leftarrow m}^f(t)$ and $c_{out(n)}^f(t) = \sum_{m\in\Omega(n)} c_{n\to m}^f(t)$. Then, queue evolution must satisfy:

$$q_n^f(t+1) \leq \max\left[q_n^f(t) - c_{out(n)}^f(t), 0\right] + c_{in(n)}^f(t) + a_n^f(t) \tag{17}$$

This cross-layer mechanism allows calculation of the optimal fair allocation vector $\overset{*}{\mathbf{x}}$, but, depending on how interference is considered, the complexity of expression (16) will vary. On this basis, we will take a closer look at the different steps to approach the solution of expression (16) under each interference model.

Under the AI model each link capacity depends on the instantaneous schedule $\boldsymbol{\pi}(t) \in \Pi$, and to exhaustively search for the best schedule (16) in each instant it is necessary to compute all the capacities in the system for all the $\mathcal{M}$ valid schedules (i.e., the total number of maximal matchings in $\mathcal{G}$), which is NP-hard and rapidly becomes intractable. Instead, we prove that stochastic approximations using the PaC [16], [28], [29] algorithm (Alg. 1) converge to a point arbitrarily close to the border of Λ, where the utility sum of all users is maximized. For this it suffices that the picking process satisfies the condition that the random schedule $\tilde{\boldsymbol{\pi}}(t)$ has a strictly nonzero probability of being the optimal schedule $\overset{*}{\boldsymbol{\pi}}(t)$. In our simulations, we found the maximum throughput in mmWave networks by running this stochastic algorithm for a sufficiently long number of frames.

Our main theoretical result is an adaptation of the traditional PaC convergence proof [16] to our hybrid directive mmWave interference AI model.

Algorithm 1 Pick And Compare

for all frame t **do**
  Pick a random feasible schedule $\tilde{\boldsymbol{\pi}}(t)$.
  for all $(n \to m) \in \tilde{\boldsymbol{\pi}}(t)$ **do**
    Compute $\tilde{c}_{n\to m}(t) = \Phi_{n\to m}(\tilde{\boldsymbol{\pi}}(t))$
  end for
  Compute the aggregate weight $\tilde{\mathbf{c}}(t)^T \overset{*}{\mathbf{w}}(t)$.
  if $\mathbf{c}(t-1)^T \overset{*}{\mathbf{w}}(t) \geq \tilde{\mathbf{c}}(t)^T \overset{*}{\mathbf{w}}(t)$ **then**
    $\boldsymbol{\pi}(t) = \boldsymbol{\pi}(t-1)$
  else
    $\boldsymbol{\pi}(t) = \tilde{\boldsymbol{\pi}}(t)$
  end if
end for

Proposition 1. As the constant V in the DCC mechanism increases, the cross-layer mechanisms described above and the PaC scheduling policy for approaching (16) achieve rates $\mathbf{x}$ that are arbitrarily close to the fair allocation $\overset{*}{\mathbf{x}}$ for all the interference models considered (WI, AI, IF).

Proof. The proof is given in the appendix. □

Our proof for PaC in the AI model also covers WI and IF because these simplified interference models can be regarded as special cases of the AI model. The WI and IF approximations are computationally simpler: the first term in the denominator of expression (37) is either removed (in IF) or kept constant (in WI), so (16) can be solved deterministically with the classic Maximum Weighted Matching (MWM) graph algorithm, with complexity $\mathcal{O}(L^3)$.

Furthermore, regarding the pseudo-wired hypothesis in [11]–[13], we remark that as spatial isolation between transmissions increases, i.e. $G(i \to m | n \to m, i \to d(i)) \to 0$, the capacity regions of WI, AI and IF converge to the third model $\Lambda_{WI} \to \Lambda_{AI} \to \Lambda_{IF}$. Hence, our analysis is perfectly valid regardless of whether or not the hypothesis is fulfilled. In practice this will depend on the quality of the MIMO hardware considered when computing the beamforming gains.

IV. Numerical Results

In this section, we use simulation to study different aspects of our mmWave multihop scheduling architecture. First, we illustrate PaC stability and convergence to optimum utility in a simple

Table I
DEFAULT NETWORK PARAMETERS

Parameter	Values
Carrier Frequency	28 GHz
System Bandwidth	1 GHz
Transmission Power	30 dBm (BS), 25 dBm (RN), 20 dBm (UE)
Noise Figure	5 dB (BS), 6 dB (RN), 7 dB (UE)
Antenna	8x8 (BS), 6x6 (RN), 4x4 (UE) $\lambda/2$ uniform planar array
Connectivity	Pathloss < 164 dB

network that is small enough to compute the optimal MWS by brute force, and compare the results. Second, we study the quality of the interference models, investigate the validity of the pseudo-wired hypothesis, and discuss the need for power control. Third, we use the model to study realistic mmWave deployments in two scenarios with different cell densities.

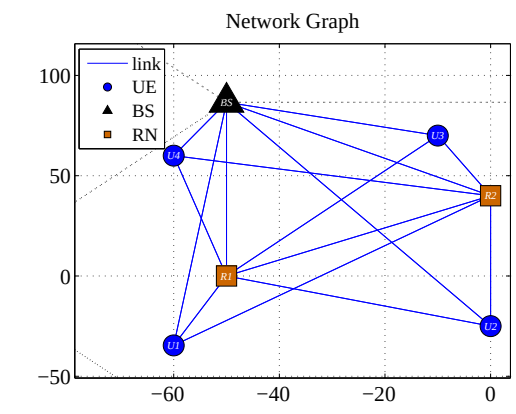
In our simulations we used the parameters in Table I, unless otherwise specified. All simulations represent outdoor urban mmWave environments.

A. Optimality Results

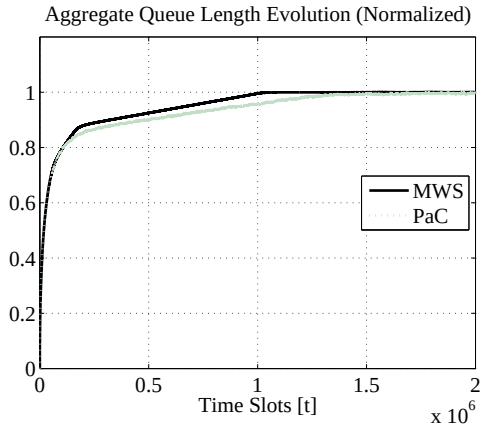
We begin by illustrating the optimality and stability of PaC with DCC in the AI model. We compared PaC with the MWS algorithm adapted to consider both the state of the queues in the network and the schedule-dependent link capacities. Since there is no MWM algorithm for this type of scenario with variable weights, we computed the adapted MWS by brute force in a simple network scenario. Figure 1 shows the scenario (Figure 1(a)) and the variations in its characteristics over time. Figure 1(b) shows the variations of the network queues, normalized by their final values when the network reaches permanent-state stable operation, and Figure 1(c) shows the growth of aggregated flow throughputs over time. This growth is essentially the same for both algorithms. Figures 1(d),1(e) and 1(f) illustrate variations in DL/UL rates and their final average values. Again the results of both algorithms matched. The oscillations in rate with PaC were due to the random nature of the algorithm.

B. Interference Results

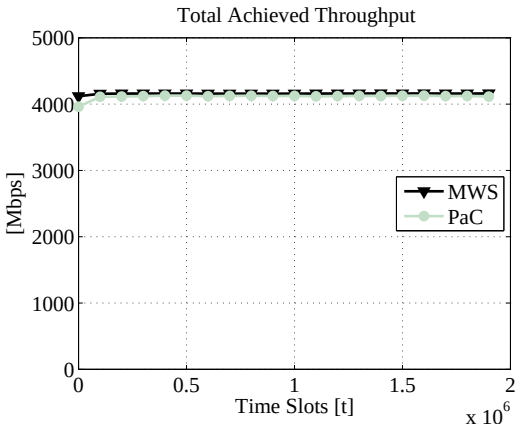
To analyze interference we tested the tightness of the IF and WI models as upper and lower bounds of realistic network capacity (given by the AI model). We considered a large realistic



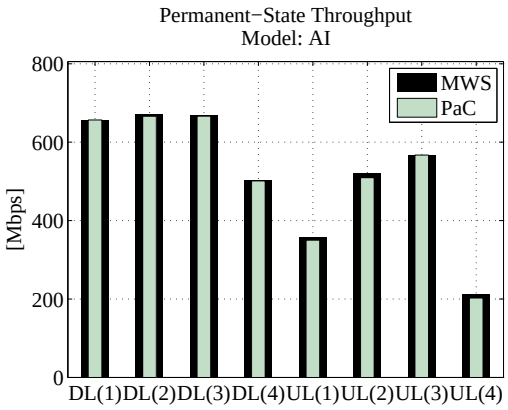
(a) Network



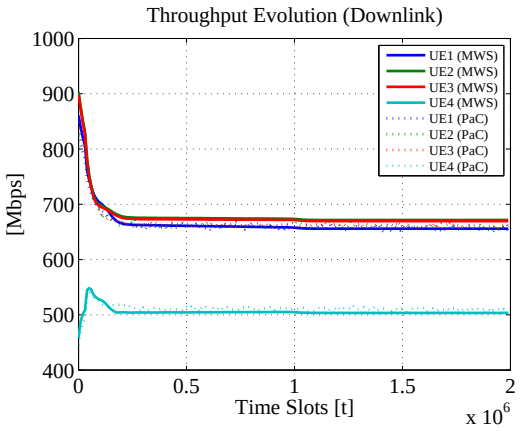
(b) Aggregated Queue Length Evolution (Normalized)



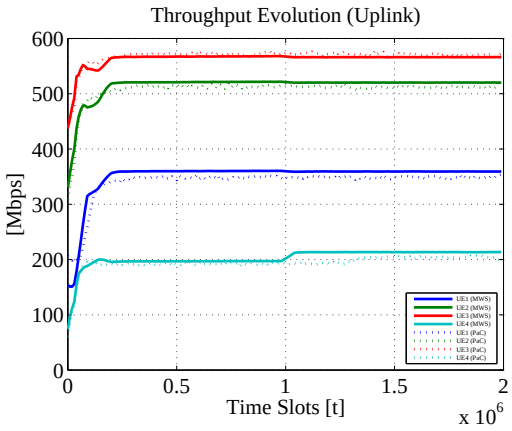
(c) Total Achieved Throughput



(d) Permanent-State Throughput



(e) DL Throughput Evolution



(f) UL Throughput Evolution

Fig. 1. Convergence and optimality in a small network.

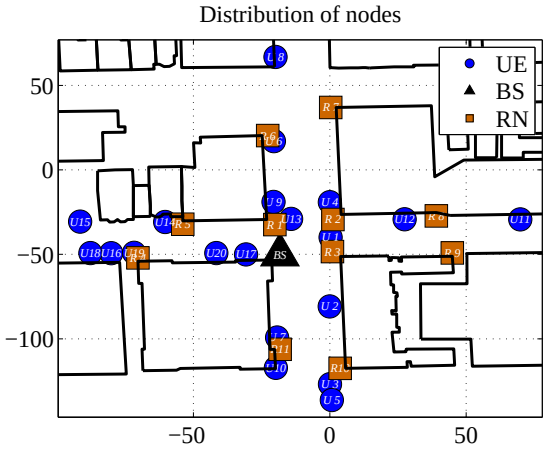
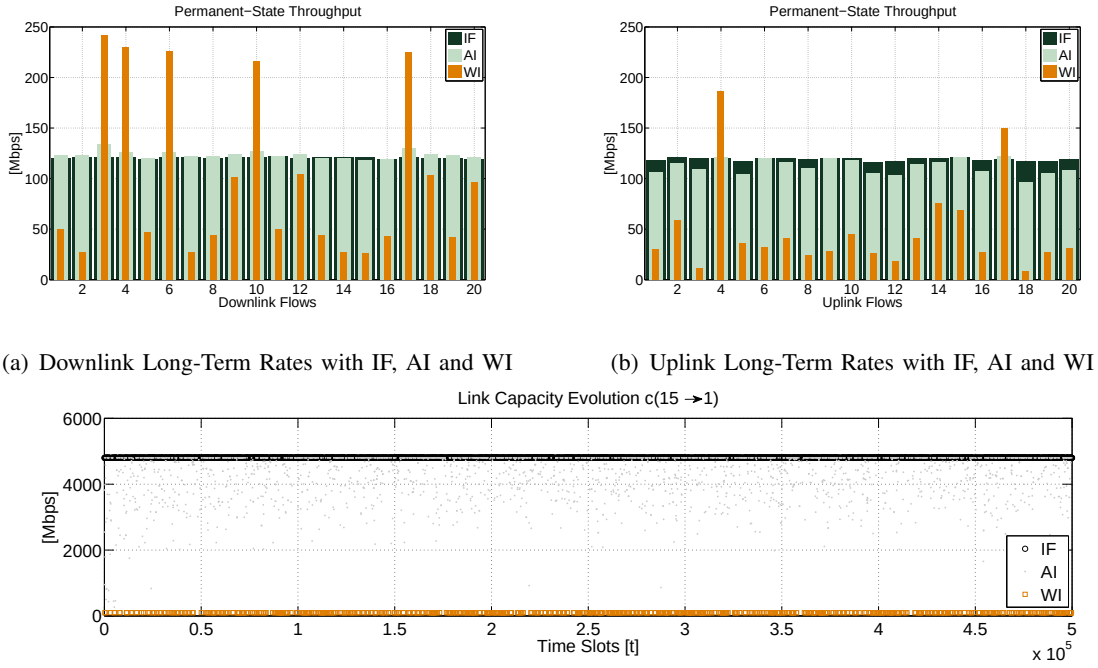


Fig. 2. Chicago city scenario

scenario with 20 UEs and 11 RNs for a 73 GHz mmWave BS picocell in a dense urban area of $160m^2$ to model mmWave interference. Figure 2 shows the urban scenario, which was made publicly available in [41], and the location of the BS, RNs and UEs. High frequency operation allows RNs and UEs to use an 8x8 $\lambda/2$ uniform planar array. The remaining parameters in Table I are unchanged. Figures 3(a),3(b) illustrates the final permanent-state rates for all users for the three interference models described in Section II-B. Interestingly, the permanent-state rates, and thus the achieved capacity (i.e. the aggregated permanent-state flow rates) for the AI and IF models, were significantly similar. We also observed a fair distribution of the resources in the AI model. Each UE has ≈ 120 Mbps at both the UL and the DL, hence $120{\cdot}2(UL/DL){\cdot}20(users) = 4800$ Mbps, which is the maximum available sum-rate in any cell, limited by the cut-set bound at the BS. As a consequence, the simplified IF model provided a tight upper bound on system performance and allowed for a realistic estimation of the achievable capacity of the mmWave system.

Additionally, the results for the WI model reveal that the pseudo-wired hypothesis does not hold in general. Taking a closer at variations in a particular link using the AI model in Figure 3(c), it is evident that some schedules notably degrade link capacity due to interference. Hence, the pseudo-wired hypothesis is highly optimistic.

The fact that interference is non-negligible is in apparent conflict with the tightness of the bound provided by the IF model. To accommodate these seemingly contradictory observations we note that the optimal long-term management of the network with the PaC algorithm accounts for



(a) Downlink Long-Term Rates with IF, AI and WI

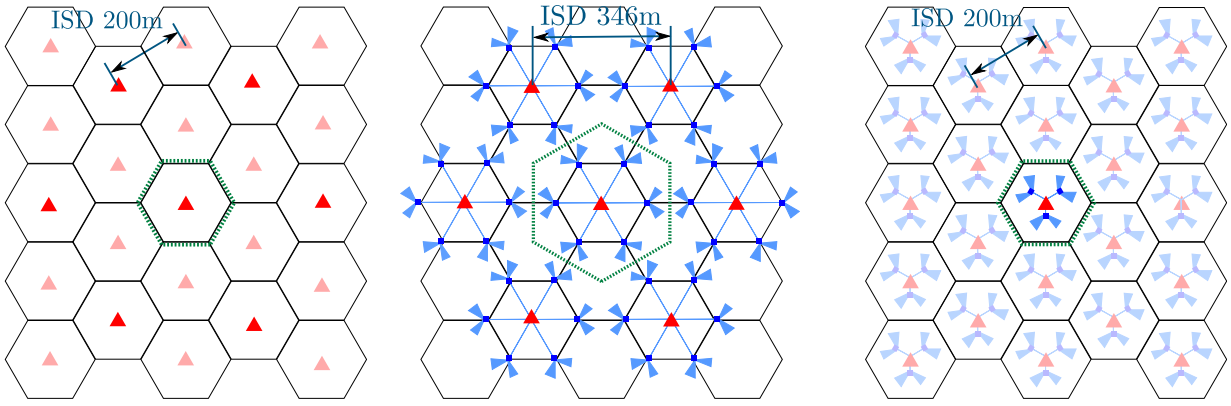
(b) Uplink Long-Term Rates with IF, AI and WI

(c) Link capacity evolution under the AI model

Fig. 3. Interference results

instantaneous link capacities in the back-pressure weights, indirectly penalizing the selection of links with high cross-interference. Therefore, this algorithm has an implicit interference avoidance effect. Moreover, its optimality guarantees that no other selection of links would result in better interference avoidance while maximizing throughput fairness among all users. In other words, interference management is as good as possible in terms of utility maximization. Furthermore, the high routing diversity of the mesh structure allows the network to route traffic around conflictive links, thereby increasing the possibilities of interference avoidance.

Since we observed that different choices of simultaneous transmitter-receiver pairs do degrade link capacity (sometimes in a very destructive way), we cannot claim that on/off power allocation is optimal with the AI model. However, on/off power allocation is optimal by definition with both the IF and WI models, and the former provides a very tight upper bound on real system performance. Therefore, it follows that any type of power control added on top of the AI model, despite improving user rates, will only achieve a minor advantage, at the cost of turning system management into a non-convex problem. Therefore, the impact of power control techniques on mmWave cellular networks will be minimal compared with their dramatic impact on the performance of traditional omnidirectional μWave cellular networks. In the extreme case when



(a) mmWave UMi network in [6] (b) Extended UMi cells for cost reduction (c) Relaying in mmWave picocells for throughput improvement

Fig. 4. Capacity evaluation scenarios.

spatial isolation of mmWave transmissions is perfect (with the introduction of denser antenna arrays) simultaneous transmitter-receiver pairs would not experience any interference, and power control would not add any value to the system in terms of throughput increase.

C. Capacity Results

We simulated two realistic mmWave cellular scenarios with multihop RNs. The reference for our comparison was the Urban Micro-cell (UMi) deployment with fully wired backhauling in [6]. This scenario is illustrated in Figure 4(a). The cells are distributed in a hexagonal grid with an Inter Station Distance (ISD) of 200m. Our first approach was to add RNs as APs with wireless backhauling to extend coverage. This scenario, represented in Figure 4(b), was designed to reduce the fiber footprint of the network by eliminating some BSs and placing RNs to extend the range of the other BSs. Our second approach consisted of enhancing capacity by adding intra-cell RNs. This scenario, represented in Figure 4(c), was designed to improve cell throughput for the same BS density as in Figure 4(a) by adding RNs halfway within the BS range. These RNs provide wireless multihop routing diversity to improve the effective SINR of cell-edge users.

Inter-cell interference is negligible because, as already noted in [6], all links in the scenario are limited by power. This means that for our purposes only intra-cell interference has to be taken into account. Therefore, in our simulations we restricted the layout to a single cell in each scenario, as delimited by the green hexagons in Figures 4(a)-4(c).

1) Coverage Extension: We simulated the scenarios delimited by a green hexagon in Figures 4(a) and 4(b) to study the value of RNs for coverage extension. In the scenario without RNs (Figure 4(a)), there is one BS every ISD $= 200$m, with 10 uniformly distributed users per BS. To construct the coverage extension scenario in Figure 4(b), we removed two of every three BSs, shown in a lighter color in Figure 4(a). The remaining BSs (one of every three), shown in red, extend their cell coverage to a radius of 200m (ISD ≈ 346.4m) and provide service to the extra users in the extended area. The total number of users is 30 per BS, and therefore user density per Km^2 remains the same. We assumed 8×8 antenna arrays at the RNs and always deployed in BS Line of Sight (LOS).

Table II
UMI COVERAGE EXTENSION COMPARISON

ISD	#RNs	A.A.^{2} (UEs)	Cell Cap		Cell Edge (P_5)		Mean		Median		Outage
			DL	UL	DL	UL	DL	UL	DL	UL	
200m	0	4x4	1688.40	1338.76	6.91	1.52	168.84	133.88	250.31	161.51	0%
346.4m	6	4x4	2470.30	1962.32	35.19	9.08	82.34	65.41	84.82	80.30	0%
200m	0	8x8	1750.28	1476.51	39.29	7.82	175.03	147.65	237.51	209.00	0%
346.4m	6	8x8	2529.26	2090.52	57.86	20.60	84.31	69.68	83.82	80.59	0%

The results of the simulation are presented in Table II. We considered the case of 4×4 and 8×8 antenna array at the UEs. The primary metrics of coverage extension are outage, which is maintained, and cell edge rates, which improve in our simulations. Furthermore, the increased number of users per cell requires a thinner splitting of cell sum-rate. It must be noted that the maximum rate of the physical layer $C_{\max}$ is 4.8 Gbps. A single BS connection with an ideal SINR operating 100% of the time in either the UL or the DL would achieve this rate. Accordingly, if we perform a cut-set bound on the sum-rate of the cell at the BS, the aggregated UL and DL sum-rate will not be able to exceed this value. To this extent, the RN scheduling scheme performs very close to the bound, with 92.35% in the case of the 4×4 antenna array and 96.25% in the case of the 8×8 array, which is more than 1 Gbps above the aggregate rate of the dense UMi deployment without RNs. However, since there are three times more users per BS, sharing this aggregate rate leads to a smaller average rate per user. A notable property of this scheme is that it distributes coverage more evenly among users, as can be observed by the improvement in "cell-edge" rates (i.e., 5th percentile or P_5).

2) Throughput Increase: We simulated the scenarios delimited by a green hexagon in Figure 4(c) to study the throughput increase using RNs for the same BS density. We departed again from the scenario without RNs, with ISD $= 200$m and 10 uniformly distributed users per cell. However, as previously mentioned, in this case we added RNs while keeping an ISD of 200 meters. RNs were distributed regularly in circumferences with a radius of 50m around the BSs, and we assumed that they were always deployed in BS LOS. We performed simulations with 2 and 4 RNs per cell.

Table III
PERFORMANCE RESULTS FOR AN URBAN MMWAVE PICOCELL

#RNs	Cell Cap		Cell Edge (P_5)		Mean		Median		% $C_{\max}$
	DL	UL	DL	UL	DL	UL	DL	UL	
0	2094.41	1894.80	10.09	3.40	209.44	189.48	292.83	290.34	83.1%
2	2369.51	2279.48	28.29	5.99	236.95	227.95	257.37	245.22	96.8%
4	2444.86	2334.11	238.56	185.25	244.49	233.41	242.98	239.70	99.6%

Table III shows the results of the simulation. The primary metric for throughput increase is aggregate rate. As we discussed above, the sum-rate of the network is physically limited by the cut set at the BS physical layer, which can process $C_{\max} = 4.8$ Gbps at most. In a UMi scenario without RNs, the cell is inefficient by nearly 1 Gbps, or approximately 20% of $C_{\max}$. With the addition of two RNs per cell, efficiency increases nearly to the maximum level and the average cell edge rates are approximately doubled. However, cell edge rates might still experience degradation if all the channel instances of certain UEs have a poor SINR. But when the number of RNs is high enough for all UEs to be served by at least one link with a high SINR, the "cell-edge" rates experience a dramatic improvement.

It is noteworthy that the DL/UL rates of the worst user in the cell are pretty close to those of a mean user within the cell. Moreover, this effect is enforced by the fair operation of our algorithm, which distributes full capacity evenly among all users in the cell. In this type of network, the traditional concept of cell-edge looses relevance, as interference is not a fundamental limitation and the bottom percentile of users is not even physically located at the border of the cells. In an ideal scenario, with interference-free links with full available capacity $C_{\max}$ (which is not the case), all users would have 240 Mbps for both the DL and the UL, which is remarkably close to the results in Table III for the 4 RNs.

V. Conclusions and Future Work

Among other technological enablers, future 5G networks will rely on mmWave bands to increase capacity. Multihop architectures will help to reduce transmission distances and achieve full exploitation of the extended spectrum. Highly directive beamforming transmitters and receivers will also be required to overcome the higher pathloss of 5G bands, with the added benefit of providing spatial isolation between transmissions. It will be possible to subsequently drop the traditional limiting global duplex coordination of BSs and UEs. Propagation studies of early 5G projects have been promising, and they are producing innovative physical layer solutions that will bring major changes to the core philosophy of cellular network operation. These solutions, in turn, will pose new research challenges for scheduling methods capable of harnessing the full potential of the network. To leverage this potential, it will be of interest to review the work on ad-hoc wireless networks, as this has traditionally dealt with multihop operation. This will be challenging, since a careful revision of the physical layer hypotheses in the field will be required to accurately address mmWave bands.

In this work we have addressed the problem of finding the maximum throughput of multihop mesh cellular mmWave networks under fairness requirements, proposing a completely dynamic duplexing resource allocation mode without UL/DL restrictions for simultaneous transmissions. We solved the problem via the well-known cross-layer decomposition framework. Although some recent works have assumed a pseudo-wired behavior for mmWave links, we studied the problem by considering real interference between simultaneous transmissions. Our proposal is a hybrid graph-oriented SINR interference model, which we call the Actual Interference (AI) model. We have demonstrated that the stochastic PaC algorithm operates the network optimally.

Furthermore, we analyzed model variants to discuss the importance of interference, power allocation, and the pseudo-wired hypothesis in previous mmWave works, by proposing upper and lower bounds on capacity derived from simplified static capacity models with topology-awareness: the SNR Interference Free model (IF), and the SINR Worst Interference model (WI).

Finally, we have provided several numerical evaluations of the analysis via simulation. Firstly, we verified the throughput-utility optimality of PaC for the realistic AI model on a network small enough for the deterministic optimal solution to be computable by brute force. Secondly, we studied the impact of schedule-dependent interference on the capacity of the network; by simulating the model with all three interference schemes (IF, AI and WI). We found that the

hypothesis of a pseudo-wired link behavior does not hold in general, but the IF upper bound provides a very tight approximation of the performance of the AI model. Our interpretation is that the existence of high spatial isolation for *some* links is sufficient for optimal scheduling to behave as an implicit interference avoidance mechanism, making interference negligible. Since power allocation mainly permits the management of interference, power control may have a much less relevant role in future mmWave multihop cellular networks. Thirdly, we simulated the capacity of mmWave cellular architectures using RNs for two different purposes: coverage extension (or backhauling cost reduction) and throughput increase. Our observations illustrate that RNs help to improve usage of the BS physical layer, as the BS can transmit at maximum spectral efficiency more often. In the coverage extension scenario, the BS sum-rate increases moderately and is distributed very efficiently by the RNs among a large number of users. In the throughput increase scenario, the BS sum-rate increases for a given coverage area and number of users, and the gains translate into an increment in average rates and cell edge rates per user.

APPENDIX

A. Throughput Optimality

The proofs in this Appendix (A and B) are closely related to the proofs in [16], yet considering soft interference. In Appendix A we prove that network queues are stable for throughput maximization with a PaC scheduling policy under the AI model. Note that this proof (like that in Appendix B) also holds for the IF and WI schemes as simplified cases. In other words, the PaC policy asymptotically achieves the same solution as MWS scheduling, as stated in Proposition 1.

Note also that the proof for MWS for the three schemes is a simplification of the proof in this Appendix involving the removal of the error term introduced by PaC due to imperfect scheduling.

Queue evolution is given by,

$$q_n^f(t+1) \leq \max\left[q_n^f(t) - c_{out(n)}^f(t), 0\right] + c_{in(n)}^f(t) + a_n^f(t) \tag{18}$$

(the inequality holds strictly when there is no data to send). To support subsequent calculations, we use the following expression:

$$(q_n^f(t+1))^2 \leq (\underbrace{\max[q_n^f(t) - c_{out(n)}^f(t), 0]}_{\max[a-b,0]=a-\min[a,b]\leq a-b, \quad a,b\geq 0} + c_{in(n)}^f(t) + a_n^f(t))^2$$

The state of the network $\mathbf{s} = (\mathbf{q}, \mathbf{c})$ forms a Markov process whose stability we want to prove. Let us define the Lyapunov function of the state $L(\mathbf{s}(t)) := \sum_{n,f}(q_n^f(t))^2$. The single-step Lyapunov mean drift function for a given $\mathbf{s} = (\mathbf{q}, \mathbf{c})$ is:

$$\Delta(t) = \mathbb{E}[L(\mathbf{s}(t+1)) - L(\mathbf{s}(t))|\mathbf{s}(t)] \leq \mathbb{E}[2\mathbf{q}^T(t)(\mathbf{c}_{in}(t) + \mathbf{a}(t) - \mathbf{c}_{out}(t)) + b_2|\mathbf{s}(t)] + b_1 \quad (19)$$

where the unfinished service is bounded by $b_1 = NF(3\Omega_{max}^2 C_{max}^2 + 4\Omega_{max}C_{max} + 2A_{max})$ and $b_2 = 2N^2C_{max}(A_{max}F + C_{max}) + A_{max}^2F^2N^2$ is a constant bound given that $\sum_{n,f} a_n^f(t) \leq A_{max}FS$ and $\sum_{n,f}\sum_{m\in\Omega(n)} c_{n\leftarrow m}^f(t) \leq C_{max}\Omega_{max}N$ (the same for $n \rightarrow m$), where S is the total number of sources, Ω_{max} is the maximum degree of $\mathcal{G}$, and the maximum capacity is bounded by $C_{\max} = \max(\mathbf{c}^{IF}$, which is the maximum interference-free link capacity across the network.

Then, in order to study the expectation in (19), we consider the following. First, we introduce the optimal scheduling term $\pm \overset{*}{\mathbf{c}}$ which accounts for potential imperfect schedules during the PaC policy. Given that $\mathbb{E}[\mathbf{a}(t)] = \boldsymbol{\lambda}$ we can write

$$\mathbf{q}(t)^T(\mathbf{c}_{in}(t)+\boldsymbol{\lambda}(t)-\mathbf{c}_{out}(t)) = \mathbf{q}(t)^T(\overset{*}{\mathbf{c}}_{in}(t)+\boldsymbol{\lambda}(t)-\overset{*}{\mathbf{c}}_{out}(t))+\mathbf{q}(t)^T(\overset{*}{\mathbf{c}}_{out}(t)-\overset{*}{\mathbf{c}}_{in}(t)+\mathbf{c}_{in}(t)-\mathbf{c}_{out}(t))$$

Now, since $\boldsymbol{\lambda} \in \Lambda_\epsilon \Rightarrow \exists\ \mathbf{c}(\Lambda_\epsilon) \setminus \boldsymbol{\lambda} = \mathbf{c}_{out}(\Lambda_\epsilon)(t) - \mathbf{c}_{in}(\Lambda_\epsilon)(t) - \epsilon\mathbf{1}$, we have

$$= \mathbf{q}(t)^T(\mathbf{c}_{in}(\Lambda_\epsilon)(t) + \boldsymbol{\lambda}(\Lambda_\epsilon)(t) - \mathbf{c}_{out}(\Lambda_\epsilon)(t)) - \epsilon\mathbf{q}(t)^T\mathbf{1} - \mathbf{q}(t)^T(\overset{*}{\mathbf{c}}_{out}(t) - \overset{*}{\mathbf{c}}_{in}(t))$$

$$+\mathbf{q}(t)^T(\overset{*}{\mathbf{c}}_{out}(t) - \overset{*}{\mathbf{c}}_{in}(t) + \mathbf{c}_{in}(t) - \mathbf{c}_{out}(t)) \leq -\epsilon\mathbf{q}(t)^T\mathbf{1} + \mathbf{q}(t)^T(\overset{*}{\mathbf{c}}_{out}(t) - \overset{*}{\mathbf{c}}_{in}(t) + \mathbf{c}_{in}(t) - \mathbf{c}_{out}(t))$$

Given that $\mathbf{1} \geq \frac{1}{C_{max}\Omega_{max}}(\overset{*}{\mathbf{c}}_{out}(t)-\overset{*}{\mathbf{c}}_{in}(t))$ and $\overset{*}{w}(t) = \mathbf{q}(t)^T\overset{*}{\mathbf{c}}(t)$, $w(t) = \mathbf{q}(t)^T\mathbf{c}(t)$, $\exists\ \mathbf{c}(\Lambda_\epsilon)\setminus\boldsymbol{\lambda} \leq -\frac{\epsilon}{C_{max}\Omega_{max}}\overset{*}{w}(t) + \overset{*}{w}(t) - w(t)$, we can rewrite expression (19) as

$$\Delta(t) \leq 2\mathbb{E}[-\frac{\epsilon}{C_{max}\Omega_{max}}\overset{*}{w}(t) + \overset{*}{w}(t) - w(t)|\mathbf{s}(t)] + \mathbb{E}[b_2|\mathbf{s}(t)] + b_1 \quad (20)$$

Now, the T-step Lyapunov mean drift is,

$$\Delta_T(t) = \mathbb{E}[L(\mathbf{s}(t+T)) - L(\mathbf{s}(t))|\mathbf{s}(t)] \leq -2\frac{\epsilon}{C_{\max}\Omega_{max}}\sum_{\tau=0}^{T-1}\mathbb{E}[\overset{*}{w}(t+\tau)|\mathbf{s}(t)]$$

$$+2\sum_{\tau=0}^{T-1}\mathbb{E}[\overset{*}{w}(t+\tau) - w(t+\tau)|\mathbf{s}(t)] + \sum_{\tau=0}^{T-1}\mathbb{E}[b_2|\mathbf{s}(t)] + b_1 \quad (21)$$

It can be shown from [16] that, when the comparison of the aggregated weights is correct and there is a strictly positive probability δ of finding the optimal schedule $\overset{*}{\boldsymbol{\pi}}(t)$ at each frame t (which is guaranteed by the on/off power allocation constraint),

$$\Delta_T(t) \leq -2T\mathbf{q}^T(t)\overset{*}{\mathbf{c}}(t)\frac{\epsilon}{C_{max}\Omega_{max}} + Tb_2 + Tb_1 \leq -b_4T\frac{\epsilon}{C_{max}\Omega_{max}}\mathbf{q}(t) + b_3 \quad (22)$$

where $b_4 > 0$ stems from the fact that $\mathbf{1} \geq \frac{1}{C_{max}\Omega_{max}} \overset{*}{\mathbf{c}}(t)$ and $b_3 = T(b_2 + b_1) > 0$. Considering Foster's criteria, expression (22) guarantees network queue stability. This means that for any inelastic $\boldsymbol{\lambda} \in \Lambda$, the PaC policy can handle the traffic in the AI model.

B. Utility Maximization

We will now show that PaC achieves an optimum solution of throughput-utility-optimality for the AI model. For this purpose we will take advantage of the previous section and start from expression (19), with $t + \tau = t_\tau$ for notational convenience.

$$\Delta_T(t) = \mathbb{E}[L(\mathbf{s}(t+T)) - L(\mathbf{s}(t))|\mathbf{s}(t)] \leq \sum_{\tau=1}^{T} \mathbb{E}[2\mathbf{q}(t)^T(\mathbf{c}_{in}(t_\tau) + \mathbf{a}(t_\tau) - \mathbf{c}_{out}(t_\tau))|\mathbf{s}(t)] + b_3 \quad (23)$$

In order to introduce flow utility, we add and subtract the term $2V \sum_{\tau=1}^{T} \mathbb{E}\left[\sum_{n,f} U_{n,f}(a_n^f(t_\tau))|\mathbf{s}(t)\right]$ in the previous expression. By reordering terms we get:

$$\begin{aligned}
\Delta_T(t) \leq\ & 2V \sum_{\tau=1}^{T} \mathbb{E}\left[\sum_{n,f} U_{n,f}(a_n^f(t+\tau))|\mathbf{s}(t)\right] \\
& - 2\sum_{\tau=1}^{T} \mathbb{E}\{\mathbb{E}[\underbrace{V \sum_{n,f} U_{n,f}(a_n^f(t_\tau)) - \mathbf{q}(t_\tau)^T\mathbf{a}(t_\tau)}_{(*1)}|\mathbf{s}(t_\tau)]|\mathbf{s}(t)\} \\
& - 2\sum_{\tau=1}^{T} \mathbb{E}[\mathbf{q}(t_\tau)^T(\mathbf{c}_{out}(t_\tau) - \mathbf{c}_{in}(t_\tau))|\mathbf{s}(t)] + b_3 \qquad (24)
\end{aligned}$$

Considering the DCC, where the injected traffic will eventually be optimum, for any fraction of the optimum we have $(*1) \geq V \sum_{n,f} U_{n,f}(\overset{*}{x}_n^f(\Lambda_\epsilon)) - \mathbf{q}(t_\tau)^T \overset{*}{\mathbf{x}}(\Lambda_\epsilon)$, and we can rewrite (24) as,

$$\begin{aligned}
\Delta_T(t) \leq\ & 2V \sum_{\tau=1}^{T} \mathbb{E}\left[\sum_{n,f} U_{n,f}(a_n^f(t_\tau))|\mathbf{s}(t)\right] - 2\sum_{\tau=1}^{T} \mathbb{E}\left\{V \sum_{n,f} U_{n,f}(\overset{*}{x}_n^f(\Lambda_\epsilon)) - \mathbf{q}(t_\tau)^T\overset{*}{\mathbf{x}}(\Lambda_\epsilon)|\mathbf{s}(t)\right\} \\
& - 2\sum_{\tau=1}^{T} \mathbb{E}[\mathbf{q}(t_\tau)^T(\mathbf{c}_{out}(t_\tau) - \mathbf{c}_{in}(t_\tau))|\mathbf{s}(t)] + b_3 = 2V \sum_{\tau=1}^{T} \mathbb{E}\left[\sum_{n,f} U_{n,f}(a_n^f(t_\tau))|\mathbf{s}(t)\right] \\
& - 2VT \sum_{n,f} U_{n,f}(\overset{*}{x}_n^f(\Lambda_\epsilon)) + \underbrace{2\sum_{\tau=1}^{T} \mathbb{E}[\mathbf{q}(t_\tau)^T(\overset{*}{\mathbf{x}}(\Lambda_\epsilon) + \mathbf{c}_{in}(t_\tau) - \mathbf{c}_{out}(t_\tau))|\mathbf{s}(t)]}_{\text{We will apply results from Appendix A here}} + b_3 \\
\leq\ & 2V \sum_{\tau=1}^{T} \mathbb{E}\left[\sum_{n,f} U_{n,f}(a_n^f(t_\tau))|\mathbf{s}(t)\right] - 2VT \sum_{n,f} U_{n,f}(\overset{*}{x}_n^f(\Lambda_\epsilon)) \\
& + 2\left[-b_4 T \left(\frac{\epsilon}{C_{max}\Omega_{max}}\right)\mathbf{q}(t) + b_3\right] + b_3 \qquad (25)
\end{aligned}$$

Now, we apply expectations to eliminate conditioning and also apply a telescoping sum of Z,

$$\mathbb{E}[L(\mathbf{s}(ZT)) - L(\mathbf{s}(0))] \leq 2V \sum_{k=1}^{Z} T\mathbb{E}\left[\sum_{n,f} U_{n,f}(a_n^f(k))\right] - 2VZT \sum_{n,f} U_{n,f}(\overset{*}{x}{}_n^f(\Lambda_\epsilon)) - 2b_4 T \left(\frac{\epsilon}{C_{max}\Omega_{max}}\right) \sum_{z=1}^{Z} \mathbb{E}[\mathbf{q}(zT)] + 3Zb_3 \tag{26}$$

The desired result for utility optimality with queue stability is now achieved given that $L(\mathbf{s}) \geq 0$ for all feasible $\mathbf{s}$. First, network queues can be bounded by rearranging terms in (26):

$$\frac{1}{Z}\sum_{z=1}^{Z} \mathbb{E}[\mathbf{q}(zT)] \leq \left(VU_{max} + \frac{3b_3}{2T} + \frac{\mathbb{E}[L(\mathbf{s}(0))]}{2ZT}\right)\left(\frac{C_{max}\Omega_{max}}{\epsilon b_4}\right) \tag{27}$$

given that $L(\mathbf{s}(ZT)) \geq 0$ and $2VZT\sum_{n,f} U_{n,f}(\overset{*}{x}{}_n^f(\Lambda_\epsilon)) \geq 0$ for all t. Furthermore, we also note that, within k frames, queues become bounded as $q_n^f(t+k) \geq q_n^f(t) + k(A_{max} + \Omega_{max}C_{max})$, so

$$\frac{1}{T}\sum_{\tau=1}^{T} \mathbb{E}[\mathbf{q}(zT+\tau)] \leq TFN(A_{max} + \Omega_{max}C_{max}) = b_5 \tag{28}$$

which, combined with (27), still satisfies network queue stability for elastic traffic:

$$\begin{aligned} \limsup_{Z\to\infty} \frac{1}{ZT}\sum_{k=1}^{ZT} \mathbb{E}[\mathbf{q}(k)] \quad &\leq \quad \limsup_{Z\to\infty} \sum_{z=1}^{Z} \mathbb{E}[\mathbf{q}(zT)] + b_5 \\ &\leq \left(VU_{max} + \frac{3b_3}{2T} + \frac{\mathbb{E}[L(\mathbf{s}(0))]}{2ZT}\right)\left(\frac{C_{max}\Omega_{max}}{\epsilon b_4}\right) + b_5 \end{aligned} \tag{29}$$

We can now demonstrate utility optimality by rearranging the terms in a different way given that $-\frac{L(\mathbf{s}(ZT))}{2ZT} \geq 0$ and $b_4(\frac{\epsilon}{C_{max}\Omega_{max}})\sum_{z=1}^{Z} \mathbb{E}[\mathbf{q}(zT)] \geq 0$ for all t:

$$\frac{1}{ZT}\sum_{k=1}^{ZT} \mathbb{E}[\sum_{n,f} U_{n,f}(a_n^f(k))] \geq \sum_{n,f} U_{n,f}(\overset{*}{x}{}_n^f(\Lambda_\epsilon)) - \frac{1}{V}\left(\frac{L(\mathbf{y}(0))}{2ZT} + \frac{3b_3}{2T}\right) \tag{30}$$

By applying Jensen's Inequality ($\mathbb{E}[U(\mathbf{X})] \leq U(\mathbb{E}[\mathbf{X}])$), we have

$$\sum_{n,f} U_{n,f}\left(\frac{1}{ZT}\sum_{k=1}^{ZT} \mathbb{E}[a_n^f(k)]\right) \geq \sum_{n,f} U_{n,f}(\overset{*}{x}{}_n^f(\Lambda_\epsilon)) - \frac{b_6}{V} \tag{31}$$

Given that $Z \to \infty$:

$$\sum_{n,f} U_{n,f}\left(x_n^f\right) \geq \sum_{n,f} U_{n,f}(\overset{*}{x}{}_n^f(\Lambda_\epsilon)) - \frac{b_6}{V} \tag{32}$$

The results in expressions (29) and (32) prove that, as constant V increases, the long-term flow rate vector $\mathbf{x}$ becomes closer to the optimum $\overset{*}{\mathbf{x}}$, while queue stability is maintained (Prop. 1).

REFERENCES

[1] Cisco, "Cisco Visual Networking Index: Forecast and Methodology, 2013–2018," *Cisco VNI Report (white paper)*, June 2014.

[2] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Communications Magazine*, vol. 49, no. 6, pp. 101–107, 2011.

[3] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.

[4] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five Disruptive Technology Directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, February 2014.

[5] S. Rangan, T. S. Rappaport, and E. Erkip, "Millimeter-Wave Cellular Wireless Networks: Potentials and Challenges," *Proceedings of the IEEE*, vol. 102, no. 3, pp. 366–385, March 2014.

[6] M. R. Akdeniz, Y. Liu, M. K. Samimi, S. Sun, S. Rangan, T. S. Rappaport, E. Erkip, and S. Member, "Millimeter Wave Channel Modeling and Cellular Capacity Evaluation," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. Jun., pp. 1164–1179, Apr. 2014.

[7] F. Gomez-Cuba, S. Rangan, and E. Erkip, "Scaling laws for infrastructure single and multihop wireless networks in wideband regimes," in *IEEE International Symposium on Information Theory (ISIT)*, July 2014.

[8] C. Hoymann, W. Chen, J. Montojo, A. Golitschek, C. Koutsimanis, and X. Shen, "Relaying operation in 3GPP LTE: challenges and solutions," *Communications Magazine, IEEE*, vol. 50, no. 2, pp. 156–162, 2012.

[9] S. Parkvall, A. Furuskar, and E. Dahlman, "Evolution of LTE toward IMT-advanced," *IEEE Communications Magazine*, vol. 49, no. 2, pp. 84–91, 2011.

[10] F. Gomez-Cuba and F. J. Gonzalez-Castaño, "Improving Third-Party Relaying for LTE-A: A Realistic Simulation Approach," in *IEEE International Conference on Communications (ICC)*, June 2014.

[11] R. Mudumbai, S. Singh, and U. Madhow, "Medium Access Control for 60 GHz Outdoor Mesh Networks with Highly Directional Links," in *IEEE INFOCOM*, April 2009, pp. 2871 – 2875.

[12] S. Singh, R. Mudumbai, and U. Madhow, "Interference Analysis for Highly Directional 60-GHz Mesh Networks: The Case for Rethinking Medium Access Control," *IEEE/ACM Transactions on Networking*, vol. 19, no. 5, pp. 1513–1527, October 2011.

[13] T. S. Rappaport, E. Ben-Dor, J. N. Murdock, and Y. Qiao, "38 GHz and 60 GHz Angle-dependent Propagation for Cellular & Peer-to-Peer Wireless Communications," in *IEEE International Conference on Communications (ICC)*, June 2012.

[14] A. Ghosh, T. A. Thomas, M. C. Cudak, R. Ratasuk, P. Moorut, F. W. Vook, T. S. Rappaport, G. MacCartney, S. Sun, and S. Nie, "Millimeter wave enhanced local area systems: A high data rate approach for future wireless networks," *Selected Areas in Communications, IEEE Journal on*, vol. 32, no. 6, pp. 1152–1163, June 2014.

[15] S. Sun, T. Rappaport, R. Heath, A. Nix, and S. Rangan, "Mimo for millimeter-wave wireless communications: beamforming, spatial multiplexing, or both?" *IEEE Communications Magazine*, no. December, pp. 110–121, 2014.

[16] A. Eryilmaz, A. Ozdaglar, D. Shah, and E. Modiano, "Distributed cross-layer algorithms for the optimal control of multihop wireless networks," *IEEE/ACM Transactions on Networking*, vol. 18, no. 2, pp. 638–651, April 2010.

[17] G. R. MacCartney and T. S. Rappaport, "73 GHz millimeter wave propagation measurements for outdoor urban mobile and backhaul communications in new york city," in *Proc. IEEE ICC*, 2014, pp. 4862–4867.

[18] W. Roh, J.-Y. Seol, J. Park, B. Lee, J. Lee, Y. Kim, J. Cho, K. Cheun, and F. Aryanfar, "Millimeter-Wave Beamforming as an Enabling Technology for 5G Cellular Communications: Theoretical Feasibility and Prototype Results," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 106–113, February 2014.

[19] S. Singh, M. N. Kulkarni, A. Ghosh, and J. G. Andrews, "Tractable model for rate in self-backhauled millimeter wave cellular networks," *arXiv preprint arXiv:1407.5537*, 2014.

[20] K. Venugopal, M. C. Valenti, and R. W. Heath Jr, "Interference in finite-sized highly dense millimeter wave networks," in *Proc. IEEE Information Theory and Applications (ITA)*, San Diego, CA, Feb. 2015.

[21] X. An and R. Hekmat, "Directional mac protocol for millimeter wave based wireless personal area networks," in *Proc. IEEE VTC*, 2008, pp. 1636–1640.

[22] J. Qiao, L. X. Cai, and X. Shen, "Multi-hop concurrent transmission in millimeter wave wpans with directional antenna," in *Proc. IEEE ICC*, 2010, pp. 1–5.

[23] C. N. Barati, S. A. Hosseini, S. Rangan, P. Liu, T. Korakis, and S. S. Panwar, "Directional Cell Search for Millimeter Wave Cellular Systems," *Arxiv preprint*, 2014. [Online]. Available: http://arxiv.org/abs/1404.5068

[24] A. Alkhateeb, O. El Ayach, G. Leus, and R. Heath, "Channel estimation and hybrid precoding for millimeter wave cellular systems," *IEEE J. Selected Topics in Signal Processing*, vol. 8, no. 5, pp. 831–846, Oct 2014.

[25] P. A. Eliasi, S. Rangan, and T. S. Rappaport, "Low-rank spatial channel estimation for millimeter wave cellular systems," *arXiv preprint arXiv:1410.4831*, 2014.

[26] F. P. Kelly, "Charging and rate control for elastic traffic," *European Transactions on Telecommunications*, vol. 8, pp. 33–37, January-February 1997.

[27] F. Kelly, A. K. Maulloo, and D. K. H. Tan, "Rate control in communication networks: Shadow prices, proportional fairness and stability," *Journal of the Operational Research Society*, vol. 49, pp. 237–252, March 1998.

[28] L. Tassiulas, "Linear complexity algorithms for maximum throughput in radio networks and input queued switches," in *IEEE INFOCOM*, vol. 2, March 1998, pp. 533–539.

[29] A. Eryilmaz, A. Ozdaglar, and E. Modiano, "Polynomial complexity algorithms for full utilization of multi-hop wireless networks," in *IEEE INFOCOM*, May 2007, pp. 499–507.

[30] Y. Yi, A. Proutière, and M. Chiang, "Complexity in wireless scheduling: Impact and tradeoffs," in *ACM International Symposium on Mobile Ad Hoc Networking and Computing*, 2008, pp. 33–42.

[31] A. Zhou, M. Liu, Z. Li, and E. Dutkiewicz, "Cross-Layer Design for Proportional Delay Differentiation and Network Utility Maximization in Multi-Hop Wireless Networks," *IEEE Transactions on Wireless Communications*, vol. 11, no. 4, pp. 1446 – 1455, April 2012.

[32] L. Lin, X. Lin, and N. B. Shroff, "Low-Complexity and Distributed Energy Minimization in Multi-hop Wireless Networks," *IEEE/ACM Transactions on Networking*, vol. 18, no. 2, pp. 501–514, April 2010.

[33] M. Neely, E. Modiano, and C. ping Li, "Fairness and optimal stochastic control for heterogeneous networks," in *IEEE INFOCOM*, vol. 3, March 2005.

[34] H.-W. Lee, E. Modiano, and L. B. Le, "Distributed Throughput Maximization in Wireless Networks via Random Power Allocation," *IEEE Transactions on Mobile Computing*, vol. 11, no. 4, April 2012.

[35] H. Ju, B. Liang, J. Li, and X. Yang, "Dynamic Power Allocation for Throughput Utility Maximization in Interference-Limited Networks," *IEEE Wireless Communications Letters*, vol. 2, no. 1, pp. 22–25, February 2013.

[36] G. D. Celik and E. Modiano, "Scheduling in Networks with Time-Varying Channels and Reconfiguration Delay," in *IEEE INFOCOM*, March 2012, pp. 990–998.

[37] X. Lin, N. B. Shroff, and R. Srikant, "A tutorial on cross-layer optimization in Wireless Networks," *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 0733-8716, pp. 1452 – 1463, Aug. 2006.

[38] M. J. Neely, *Stochastic Network Optimization with Application to Communication and Queueing Systems*. Morgan & Claypool, 2010.

[39] J. Kim and A. F. Molisch, "Quality-Aware Millimeter-Wave Device-to-Device Multi-Hop Routing for 5G Cellular Networks," in *IEEE International Conference on Communications (ICC)*, 2014.

[40] A. Eryilmaz, *PhD. Efficient and Fair Scheduling for Wireless Networks*. University of Illinois, 2005.

[41] T. Bai and R. W. H. Jr., "Coverage and rate analysis for millimeter wave cellular networks," *CoRR*, vol. abs/1402.6430, 2014. [Online]. Available: http://arxiv.org/abs/1402.6430